Noname manuscript No.
(will be inserted by the editor)

Detecting Dressing Failures using Temporal-Relational Visual Grammars

Elias Ruiz · Venet Osmani · L. Enrique Sucar · Oscar Mayora

the date of receipt and acceptance should be inserted later

Abstract Evaluation of dressing activities is essential in the assessment of the performance of patients with psycho-motor impairments. However, the current practice of monitoring dressing activity (performed by the patients in front of the therapist) has a number of disadvantages when considering the personal nature of dressing activity as well as inconsistencies between the recorded performance of the activity and performance of the same activity carried out in the patients' natural environment, such as their home. As such, a system that can evaluate dressing activities automatically and objectively would alleviate some of these issues. However, a number of challenges arise, including difficulties in correctly identifying garments, their position in the body (partially of fully worn) and their position in relation to other garments. To address these challenges, we have developed a novel method based on visual grammars to automatically detect dressing failures and explain the type of failure. Our method is based on the analysis of image sequences of dressing activities and only requires availability of a video recording device. The analysis relies on a novel technique which we call *temporal-relational visual grammar*; it can reliably recognize temporal dressing failures, while also detecting spatial and relational failures. Our method achieves 91% precision in detecting dressing failures performed by 11 subjects. We explain these results and discuss the challenges encountered during this work.

Keywords Assessing Dressing Activity · Pervasive Healthcare · Spatial Relationships · Structural Pattern Recognition · Temporal Grammars · Visual Grammars

1 Introduction

Dressing activity is a complex skill that is taken for granted in able-bodied and able-minded individuals. However, following cognitive and motor impairments, this self-care task can become very problematic, considering that 54% of stroke survivors are unable to dress independently after six

Elias Ruiz
Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro No. 1, Sta. María Tonanzintla, Puebla, México C.P. 72840

Venet Osmani
Fondazione Bruno Kessler (FBK), Trento, TN, 38123, Italy

L. Enrique Sucar
Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro No. 1, Sta. María Tonanzintla, Puebla, México C.P. 72840

Oscar Mayora
Fondazione Bruno Kessler (FBK), Trento, TN, 38123, Italy

months (Edmans and Lincoln 1990) and 36% after two years (Edmans et al 1991). While there is clinical evidence to suggest that dressing practice, provided by occupational therapists, can be beneficial (Walker et al 1996), there is very little prior work (Matic et al 2012) in using technology to automatically monitor dressing activities and report the different types of failures during dressing activities. The work in this paper aims to address this gap in the research literature by investigating the feasibility of a computer vision based system, using a novel type of visual grammar, which we call *temporal-relational visual grammar*, to automatically monitor and detect different types of failures in dressing activities. The choice of using computer vision is based on the fact that such system is inexpensive and already present in many homes, for example built-in cameras and web-cams found in personal computers. In addition, dressing activity images are processed on the device and failures communicated in situ, without being transmitted outside patient's home, while only relevant parts of the images are used (for example the face is automatically blurred) preserving patients' privacy. Lastly, our system does not require modification or tagging of garments and is fully reliant on image processing and recognition based on temporal-relational visual grammars.

Our previous work in this area, (Matic et al 2010, 2012), relied on manual tagging of items of clothing with RFID tags, in combination with a computer vision based system, to automatically detect dressing failures. In this paper, we build upon our previous work by eliminating the need for manual RFID tagging of clothing items, relying solely on computer vision and a temporal-relational visual grammar to automatically detect dressing failures.

In order to investigate the feasibility of using temporal-relational visual grammars to detect dressing failures, we have recruited eleven test subjects, not connected with this research. After agreeing to informed consent, each subject was asked to perform the dressing task by choosing any combination of clothing items, without assistance. Dressing activity was carried out in a dressing room, as shown in Fig 1, where a video recording camera recorded each subject. Initially, test subjects performed the correct dressing task and then they were free to choose from a set of dressing failures identified from current research literature (Sunderland et al 2006; Walker and Lincoln 1991). We have analysed three types of failures, namely:

i) *temporal*: where the sequence of garments is incorrect (for example a shirt is put on after a jacket);

ii) *relational*: where the garments are put on incorrectly in relation to body (for example a jacket is put on backwards); and,

iii) *spatial*: where the garments are put on partially (for example only one sleeve of a jacket is put on).

Our results show that we can reliably identify temporal dressing failures, while it is more challenging to automatically identify relational and spatial failures. Without considering failure type, we can detect dressing failures with 91% precision, which may be useful as an indicator of disease progression or improvement of patients' state.

Our contribution is twofold: (i) this is the first work to investigate automatic detection of dressing failures relying solely on visual information obtained from a single camera; and (ii) we develop a novel extension of symbol-relational grammars, which we call temporal-relational visual grammars. Based on this representation we can encode rules for correct dressing and various failure types. In addition, we combine an image processing and classification component with a rule-based parsing algorithm to detect and explain the failures in dressing activities.

The rest of the paper is organised as follows: Section 2 summarizes related work in monitoring dressing activity and related work on visual grammars for object detection. Section 3 provides an overview of visual grammars. In section 4 we describe our methodology, while Section 5 presents the experimental results. Section 6 summarises the work and outlines our future research plans.

Fig. 1 Three types of failures were considered in the dressing activities. The proposed model can detect the four possible cases in controlled environments: a) correct dressing example b) temporal failure– wrong order garment, c) spatial failure– partially worn garment, and d) relational failure– backwards.

2 Related Work

2.1 Monitoring dressing activity in patients

While there is clinical evidence to suggest that dressing practice provided by occupational therapists can be beneficial (Walker et al 1996), there is very little prior work specifically focused on dressing activities. Bahle et al (2014) monitor hospital activities using a smartphone carried out by nurses. They provide results pertaining dressing activities conducted by nurses with the patients, however reporting types of dressing failures is not in their focus. Similarly, Chen et al (2012) investigate the use of location in recognising daily activities, including dressing, while the mental state of patients is recognised in Osmani (2015). However, authors do not specifically focus on dressing activity, thus do not provide any results regarding types of dressing activity failures. The challenges in recognising dressing activity are well highlighted by Chernbumroong et al (2013), where out of nine ADLs recognised, dressing was the most challenging, contributing most misclassification errors. In addition, a recent survey of visual detection of human activities (Afsar et al 2015) has found that very little attention is given to the activity of dressing.

Clinical practice of dressing assessment involves therapists periodically taking notes while the patient performs the dressing steps (Feyereisen 1999; Namazi and Johnson 1992), using the Nottingham Stroke Dressing Assessment (NSDA) (Walker and Lincoln 1991) scale, for instance. However, this approach has three considerable disadvantages: i) dressing is a personal and private activity and carrying it out in front of another person is often uncomfortable and unpleasant; ii) note taking is not only error prone, but also subjective, making it difficult to compare notes when different therapists assist the same patient. In this regard, a literature review (Walker and Walker 2001) and survey of occupational therapy dressing practices in the UK documented that therapists did not use standardised dressing assessments to evaluate dressing performance (Walker et al 2003); and iii) the presence of therapists can result in inconsistencies between the recorded performance of the activity and performance of the same activity carried out in the patients' usual environment, such as their home. This is because patients and especially the elderly will invest extra effort to carry out the activity correctly and thus vindicate their independence, as was demonstrated in a study by Brown et al (1996).

2.2 Visual grammars for object detection

There are several models that combine a visual grammar with object recognition. One of them is Qi et al (2017), where they propose an *and-or graph* to segment and predict a number of human activities. The representation of the graph is not defined in the formalism of a grammar. More explicitly, the graph operates as a spatial and temporal grammar, however this representation is limited and does not have the full potential of a visual grammar. For visual grammars there are different approaches, principally focusing on the inner structure of the object for object detection tasks, where several approaches disregard grammar representation. In Wu et al (2010); Zhu and Mumford (2006), an And-Or graph scheme to represent visual objects is used, while in Girshick et al (2011) an acyclic grammar is used to score pedestrian detection. Zhu et al (2009) propose a combination between probabilistic context free grammars and Markov Random Fields to recognise an object. In Foncubierta-Rodríguez et al (2017), they adapt a language grammar with the bag-of-visual-words paradigm for image understanding tasks whereas Friedman and Ron (2017) apply a visual grammar to social media analysis in elections. We have found a number of works related to temporal representation and analysis. For example Maio et al (2017b), outlines a method to analyse tweets with temporal and semantic relations, while a ranking method is presented in Maio et al (2017a).

The majority of previous work convey a grammar designed for a specific task, and in particular they do not consider a knowledge representation that combines spatial and temporal aspects within the grammar.

3 Background

We have chosen to use visual grammars considering their advantages over other methods, including: i) codifying rules of correct dressing and dressing failures with little effort; and ii) representation of spatial and temporal information, in addition to relational information - the core aspect for automatically detecting dressing failures. Below, we briefly explain visual grammars and proceed with the formalism of our proposed method - *temporal-relational visual grammar*.

3.1 Visual Grammars

Visual grammars (Gottfried 2015; Lakin 1987; Leborg 2006) are a way to express the knowledge observed in a visual schema using only predicates. One simple example is *person = Above(head, body)*. This predicate subsumes two parts (head and body) into a new word (person). In the world of predicates, we do not need the graphical representation: the grammar retrieves the visual information using symbols (like head or body) and relationships (Above). Using grammars provides the following advantages: The predicates are both machine and human readable, allowing interpretability of the model in almost every stage: describing the grammar, parsing the grammar in an example, understanding the relationship between the grammar and answering a query (the inference engine). A query is a question whether an object can be generated by the grammar or not. In this sense grammars are not black boxes, as opposed to other approaches that describe the world in terms of numeric features only.

Therefore, the grammars can be easily edited, making it simple to add additional knowledge to the system. This is in contrast to other methods where the implications of changing particular parameters are not easily understood. The proposed model is focused on describing the spatial and temporal relationships between garments, required to recognise failures. Considering that we need to manually describe what constitutes correct dressing and what constitutes a dressing failure, visual grammars are a suitable option to represent the knowledge of the correct sequence of garments and their position on the body.

3.2 Symbol-Relational Grammars

Transformational grammars (Chomsky 2002) are grammars where a complex element is hierarchically decomposed into simpler ones. No relationships are provided or explained. An example of a transformational grammar can be that A is transformed in bc: $A \rightarrow bc$. One can suppose that A is a train and b,c are two wagons where b is placed left from the wagon c. This decomposition can be context free or with added restrictions, but always operated in a sequence (one dimensional, such as in a line). In other words, b always has a "left" relationship with the c element. In comparison, **relational grammars** (Wittenburg and Weitzman 1996) include other dimensions by adding relationships between the elements of the grammar: the result is that relational grammars operate in a two-dimensional space (an unlimited layout, instead of a line). If we use the same example, the extension to symbol-relational grammars means that objects b and c can hold more relationships, such as *above*, *within*, *behind* and so on. Relational grammars are supported by predicate logic in order to have a richer representation including variables (such as the sky appears above everything: $above(sky, X)$).

In this paper we propose an extension of a symbol-relational (SR) grammar (Ferrucci et al 1996) because current SR-grammars do not provide an explicit way to codify temporal relationships (Allen 1983), an essential requirement for our challenge. We incorporate temporal relationships within symbol-relational grammars, in addition to spatial relationships. The inclusion of time allows us to handle temporal relationships (such as sequence of garments for example), and the detection and explanation of temporal errors becomes straightforward using a rule-based inference engine, which we have also developed. We name our proposed extension: "temporal-relational visual grammar". Previous work on visual grammars (Costagliola et al 2002; Kong et al 2006; Lakin 1987; Marriott and Meyer 1996; Mjolsness 1991) did not consider temporal relations, as those works were focused on single images; the proposed extension opens the door for future applications in image sequences or video, where temporal aspects are essential.

We now briefly describe the formalism of Symbol-Relational (SR) grammars including several examples followed by a description of the inclusion of temporal relationships.

Formally, an SR grammar is a tuple $\mathcal{G} = (V_N, V_T, V_R, S, P, R)$, where:

- V_N is a finite set of non terminal symbols.
- V_T is a finite set of terminal symbols.
- V_R is a finite set of relational symbols between $V_N \cup V_T$.
- $S \in V_N$ is the starting symbol.
- P is a finite set of labelled rules, called s-item productions of the form:

$$l : Y^0 \rightarrow \langle \mathbf{M}, \mathbf{R} \rangle$$

  where:
  - l is an integer labelling the s-production.
  - $\langle \mathbf{M}, \mathbf{R} \rangle$ is a sentence on V_R and $V_N \cup V_T$
    - $\mathbf{M}$ is a set of s-items (v, i) with $v \in V_N \cup V_T$ and i is a natural number used to distinguish different occurrences of the same symbol.
    - $\mathbf{R}$ is a set of r-items of the form $r(X^i, Y^j)$, with X^i, $Y^j \in \mathbf{M}$ and $r \in V_R$
  - $Y \in V_N$, $Y^0 \notin \mathbf{M}$
- R is a finite set of rewriting rules called r-item productions. Since we do not use this kind of productions in our model, we will omit its definition. See Ferrucci et al (1996) for details. In all cases we define R as $\varnothing$.

Conventionally, the index “0” will only be used to denote the symbol on the left-hand side of every s-production. In the right-hand side indices “2”, “3”, ... are used to express different instances of the same symbol. Index “1” is not used.

Initially we provide a number of examples using SR-grammars without temporal relations, so as to gain a better understanding how an SR grammar addresses spatial relations in an image. The following definition using a SR-grammar describes a person wearing a sweater or a shirt with jeans:

$\mathcal{G} = (V_N, \{sweater, shirt, jeans\}, \{above\}, A, P, \varnothing)$
where P is given by:

$$A^0 \rightarrow \left\langle \left\{ sweater^2, jeans^2 \right\}, \left\{ above(sweater^2,\ jeans^2) \right\} \right\rangle$$
$$A^0 \rightarrow \left\langle \left\{ shirt^2,\ jeans^2 \right\}, \left\{ above(shirt^2,\ jeans^2) \right\} \right\rangle$$

where the superscripts are used in cases where there are two or more objects of the same type; for example, if we have two sweaters of the same type, one of them is referred to with the two-superscript, the other is described with the three-superscript. It should be noted that for our application there were no two instances of the same garment and as such superscripts can be omitted.

The same symbol in the left-side of each s-production (in the example above, A^0) signifies an *Or-rule*: a person can wear a sweater above his or her jeans *or* the same person can wear a shirt *instead of* the sweater.

The detection of the visual objects in the images (for instance, the jeans or the sweater) is addressed through a classification algorithm described in the Section 4.

3.3 Temporal-Relational Visual Grammars

Several changes in SR grammars are required in order to describe the formalism of temporal-relational visual grammars. Our decision to include temporal relationships between objects stems from the need to describe temporal relationships between garments. Therefore, the definition for the Temporal-Relational Visual grammars (or TR-visual grammars, for short) is:

A TR-visual grammar is a tuple $\mathcal{G}_T = (V_N, V_T, V_R, S, P, R)$. The definition of G_T is similar to the previous for SR-grammars; the TR-visual grammars include all its temporal relationships in V_R. However, TR-visual grammars have a different formalism for the production rules. In this sense, P is a finite set of labelled rules, called s-item productions (symbols production) of the form:

$$l : Y^0 \rightarrow \langle \mathbf{M}, \mathbf{R} \rangle$$

where:

- l is an integer labelling the s-item production.
- $\langle \mathbf{M}, \mathbf{R} \rangle$ is a sentence on V_R and $V_N \cup V_T$
  - $\mathbf{M}$ is a set of s-items (v, t, i) with $v \in V_N \cup V_T$, t is a natural number to describe the frame where the symbol belongs and i is a natural number used to distinguish different occurrences of the same symbol in the same frame.
  - $\mathbf{R}$ is a set of r-items of the form $r(X_m^i, Y_n^j)$, with $X_m^i,\ Y_n^j \in \mathbf{M}$ and $r \in V_R$
- $Y \in V_N$, $Y_0 \notin \mathbf{M}$

Note, there are two associated indices for each symbol. In other words, since the superscripts are used to define instances of the same symbol in the $V_T \cup V_N$ set, we added subscripts to describe the $timeframe$ where the symbol is placed. Since in our application we do not require superscripts, we will omit them for the rest of the paper. For example: $A^0 \rightarrow Next(Shirt_1^2, Jacket_2^2)$ will be written as: $A \rightarrow Next(Shirt_1, Jacket_2)$, where the shirt belongs to the first frame and jacket belongs to the second frame. The addition of a temporal relationship can be combined with spatial relations in the same rule, however for our purpose temporal relationships will be considered in different rules. Rewriting rules will not be used, thus $R = \varnothing$. We use $Or - rules$ to explain the steps in several dressing activities; that is, the rules have the same meaning as in natural language, for example: after

a shirt, a jacket *or* a sweater can be worn. The additional index allows us to handle the temporal relationships separately. The composition can be operated at the terminal level or in meta-rules. For our purposes we perform temporal composition at terminal levels. A grammar $\mathcal{G}_T$ always comprises a complete and correct[1] dressing activity. For instance:

$G = (\{Seq,\ First\,, Second\}, \{tshirt,\ poloshirt,\ jeans\},$
$\{above,\ aligned, Next\},\ Seq,\ S,\ \varnothing).$

where S is given by the following production rules:

$$1 : First \rightarrow \langle \{tshirt_1, jeans_1\}, \{above(tshirt_1, jeans_1), aligned(tshirt_1, jeans_1)\}\rangle$$
$$2 : Second \rightarrow \langle \{poloshirt_2, jeans_2\}, \{above(poloshirt_2, jeans_2), aligned(poloshirt_2, jeans_2)\}\rangle$$
$$3 : Seq \rightarrow \langle \{First_1, Second_2\}, \{Next(First_1, Second_2)\}\rangle$$

where all the subscripts are defined according to our formalism. As we explained before, superscripts are not necessary because in our examples we don't have two or more instances of a certain symbol (garment) in a frame. The two instances of jeans are in two different frames so they are considered different objects. It should be noted that this form creates meta-rules in a hierarchical way. In order to obtain a better explanation of the transitions between each garment (instead of the frames), we decided to rewrite the previous production rules of the grammar with more detail; thus we reformulate the production rules in this way:

$$1 : First \rightarrow \langle \{tshirt_1, jeans_1\}, \{above(tshirt_1, jeans_1), aligned(tshirt_1, jeans_1)\}\rangle$$
$$2 : Second \rightarrow \langle \{poloshirt_2, jeans_2\}, \{above(poloshirt_2, jeans_2), aligned(poloshirt_2, jeans_2)\}\rangle$$
$$3 : UpperT \rightarrow \langle \{tshirt_1, poloshirt_2\},\ \{Next(tshirt_1, poloshirt_2)\}\rangle$$
$$4 : LowerT \rightarrow \langle \{jeans_1, jeans_2\},\ \{Next(jeans_1, jeans_2)\}\rangle$$
$$5 : Seq \rightarrow \langle \{First_1, Second_2, UpperT_*, LowerT_*\}, \varnothing\rangle$$

where $UpperT$ and $LowerT \in V_N$. With these new rules the explanation of the transitions is more clear than with the previous rules. $UpperT$ and $LowerT$ are non terminal elements operating between two frames; we use the star symbol in the subscript instead of the frames where they appear. We do not need to explain more spatial or temporal relationships in the last rule as the sequence is defined with the set of the non-terminal elements included in rule five. This grammar example corresponds to Fig. 6, below, an example of correct dressing.

4 Methods

A general schema of the proposed method has been outlined in the Figs. 2 – training, and 3 – parsing; composed of the following steps:

1. Train the visual garment detectors.
2. Build a model that includes the knowledge about spatial and temporal relationships, for example: "garment a appears above and is aligned with respect to garment b", where a and b are types of garments learned in the previous step.
3. Build a grammar that explains all the correct instances of a dressing activity (dressing failures are treated as any combination of garments that were not explicitly written in the production rules of the grammar).

[1] The grammar does not contain rules describing failures, meaning that if a configuration of a dressing activity cannot be explained by the grammar it is marked as a failure.

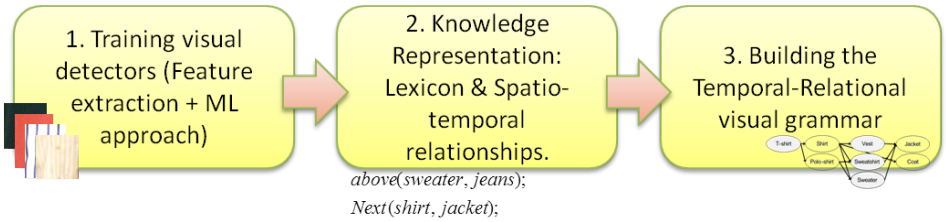

Fig. 2 Schema of the training phase of our method. It consists of three main blocks: In the first stage it extracts visual features and trains the garment detectors with a machine learning approach. In the second stage we represent the garment symbols and the spatial and temporal relationships required. Finally, we build (manually) the grammar for the dressing activities off-line.

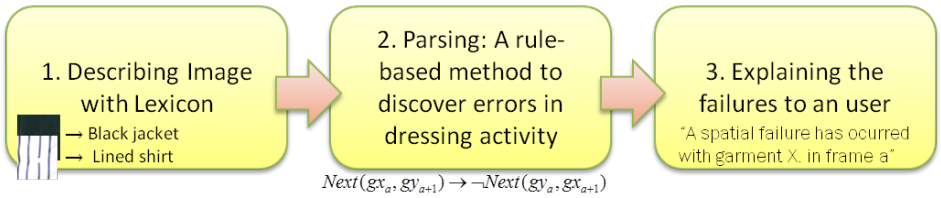

Fig. 3 Schema of the parsing phase. In the first stage each image that will be parsed is described in terms of the garment lexicon. In the second phase a rule-based method discovers errors in a sequence of images. Finally we transform the answer of the system in a sentence for the user.

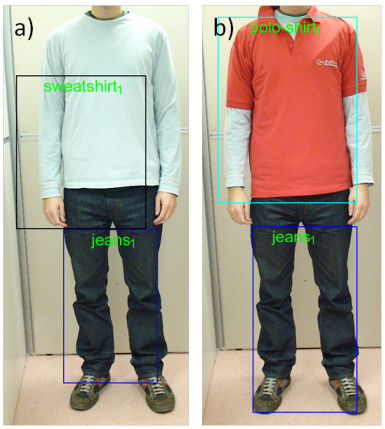

Fig. 4 Example of wrong order failure. Sometimes our model observed the garments from the previous image (long sleeves in image b) and it generated an additional error. See text for details.

4. Process a sequence of images combining the garment detectors to describe the images in terms of the garment lexicon. Then, use the rule-based method to decide if the sequence corresponds to an instance of correct dressing or to a failure. Our method explains what kind of dressing failure has been detected. For example: "For Person p_a the detected failure is wrong order of garments, since garment G_a appears before garment G_b". G_a is a sweatshirt and G_b is a polo-shirt. Another example is: "For Person p_a a spatial failure is detected, since garment G_a is partially worn, considering the model still detects the previous garment".

Examples of these failures are visually shown in the Figs 4 and 5. We now explain our method in more detail.

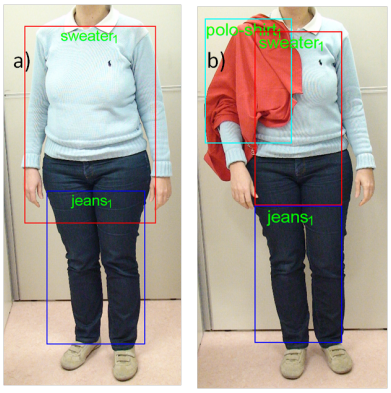

Fig. 5 Example of partial garment failure. In the experiments our grammar explained two associated errors (partial dressing and temporal order). See text for details.

4.1 Step 1: Training Visual Garment Detectors

The majority of computational models for object recognition are based on local features (Bay et al 2008; Lowe 2004; Rublee et al 2011), and/or a combination of shape, texture, edges or global features applied over patches (Dalal and Triggs 2005; Mikolajczyk and Schmid 2005). Our main goal is to recognise an specific garment. As such, we use colour histograms and texture information since in a cross-validation study these features performed better than features such as dense-Sift or Gabor filters.

We extracted colour information using colour histograms over RGB, HSV and Lab at 16 bins; and texture features using gray-level co-occurrence matrix and local binary patterns (Haralick et al 1973; Ojala et al 1996; Vedaldi and Fulkerson 2010). To obtain the previous features, we considered a simple window based approach, using a grid of patches over the image (we used patches with 70 pixels). In order to learn visual classifiers to detect the different types of garments we used Support Vector Machines (SVM) (Cortes and Vapnik 1995) with linear kernel as classifiers. This was because linear kernels performed best in cross-validation tests. In our experiments we considered 38 different garments that were part of the dataset. It should be noted that adding additional garments is straightforward since we only need to provide a visual example associated to the type of garment (for example, a blue-squared-shirt image with the 'shirt' label, a brown-lined-trousers with the 'trousers' label and so on) and this is performed only once.

We trained separate classifiers, where each classifier recognises one kind of garment, following one vs. the rest method. We used a supervised schema since at this stage we are interested in cases where we already know all the kinds of garments worn by the subjects.

We address the classifier errors through a number of strategies, namely: i) background subtraction is performed using empty background images provided by the dataset; ii) non-maxima suppression to remove false positives in the image when the classification score is low, while preserving the garment with a high score; and, iii) fusion of small patches when they correspond to the same classifier, since we do not expect a person holding two different garments of the same type (such as two sweaters or two shirts). An example of region detection is shown in Fig. 6.

4.2 Step 2: Knowledge Representation

Dressing failure detection relies on an analysis of a sequence of images where, for example, in the first frame the subject has a shirt with jeans, while in the second frame, the same subject has a jacket with the same jeans as shown in Fig 7. Thus, using spatial and temporal relationships, we can

Fig. 6 Example of correct dressing. A shirt is put it on before a polo shirt. The example has not failures in order, partial dressing or backwards.

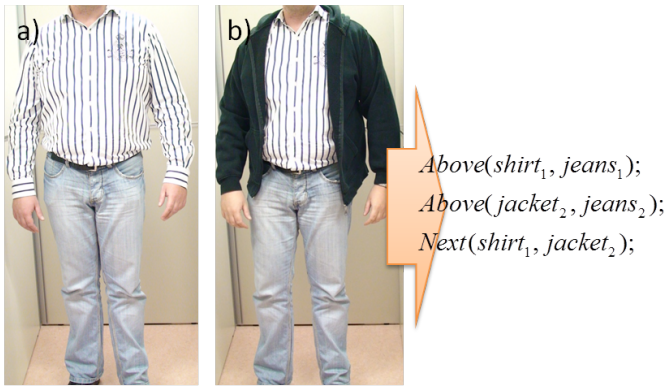

Fig. 7 Example of a correct dressing activity and the corresponding spatial and temporal relations.

write: $Above(shirt_1, jeans_1)$, $Next(shirt_1, jacket_2)$ and $Above(jacket_2, jeans_2)$. In these predicate examples we have added frame information using subscripts, which allows us to distinguish the same garment in different frames.

Each detected garment is described in terms of its spatial position with respect to the image, that is if the garment is placed in the upper or lower part of the body. This is achieved using arity-one predicates (only one argument). In this sense, $Isupper(Shirt_a)$ is an unary predicate which describes the position of the centre of the blob ($Shirt_a$) in the image. For clarification purposes, a is the number of the frame, $Shirt$ is a name for the garment and $Isupper$ is the name of the predicate.

It is important to describe what kind of spatial relationships are present in the image, for example the rule $Above(coat_2, trousers_2)$ refers to a spatial relationships between two garments in the second frame of a sequence. For transitions, we use one kind of temporal relationship: the substitution of a garment in the subsequent frame, which we call $Next$ relationship. Other kinds of temporal relationships were omitted, as they were not required. For an overview of temporal relationships see Allen (1983). The $Next$ relationship has the form: $Next(A_f, B_{f+1})$, where A and B are garments placed in two consecutive frames. We define this relationship once the following intersection is satisfied: $(A_f \cap B_{f+1})/(A_f \cup B_{f+1}) \geq \epsilon$, where ϵ is fixed to 0.5. The intersection is given by the positioning of the images in the sequence. This step (building the knowledge base) is performed automatically by

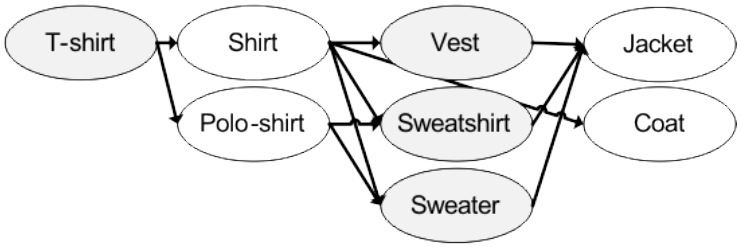

Fig. 8 An example of graphical representation of the garment order for our model. The order of the upper garments in the examples considered in the dataset is summarized in this graph. The graph for lower garments is simpler since for our database we expect no changes of the jeans/trousers during the dressing process.

our model since it only needs the sequence of images and the garments obtained in the previous step, without further intervention.

4.3 Step 3: Building the Temporal-Relational Grammar

Even if Symbol-Relational grammars can express the inclusion of temporal relationships, we hypothesise that our special treatment of temporal relationships is better since it is more explicit: we can define the frame or position in time of each symbol/object. Moreover, the proposed grammar distinguishes between temporal and spatial relationships. In this regard, the rules that we want to include are related to correct dressing activities only. If a dressing activity follows the grammar, the dressing activity is correct; if it fails, there is an error in the activity. In this manner, we are more interested in the discriminating power of the grammar, rather than creating a language of all the accepted combination of garments. To describe whether a sequence of garments in a dressing sequence of a person is accepted by the grammar, we define a grammar that comprises all the "correct dressing activities" using Or-rules for spatial and temporal relationships. Examples of temporal relationships in the grammar are:

$$
\begin{aligned}
&G \rightarrow \langle \{tshirt_a,\, shirt_{a+1}\}, \{Next(tshirt_a,\, shirt_{a+1})\}\rangle,\\
&G \rightarrow \langle \{shirt_a,\, vest_{a+1}\}, \{Next(shirt_a,\, vest_{a+1})\}\rangle,\\
&G \rightarrow \langle \{shirt_a,\, sweater_{a+1}\}, \{Next(shirt_a,\, sweater_{a+1})\}\rangle,\\
&G \rightarrow \langle \{poloshirt_a,\, jacket_{a+1}\}, \{Next(poloshirt_a,\, jacket_{a+1})\}\rangle,\\
&G \rightarrow \langle \{tshirt_a,\, poloshirt_{a+1}\}, \{Next(tshirt_a,\, poloshirt_{a+1})\}\rangle,\\
&G \rightarrow \langle \{shirt_a,\, sweatshirt_{a+1}\}, \{Next(shirt_a,\, sweatshirt_{a+1})\}\rangle,
\end{aligned}
$$

where the subscripts a, $a+1$ mean garments of two subsequent frames. *Or-rules* are obtained directly by using the same non-terminal element in the left side of the grammar.

Using a graphical representation of the predicates, Fig. 8 illustrates the correct order of upper garments from our dataset. This order was built manually, given that a person can easily describe the right order of the garments (e.g. typically a jacket or a coat should be put on last, after all the other garments). This description is a graphical representation of a common garment order. A general representation of temporal order is represented in the grammar by a set of predicates (where in the figure we only show the upper garments). The elements in this graph (grammar) can be easily extended to include other types of garments.

In the same manner, we write all the correct spatial dressing examples. Several examples are given below:

$$G \rightarrow \langle \{tshirt_a,\, trousers_a\}, \{Above(tshirt_a,\, trousers_a)\} \rangle ,$$
$$G \rightarrow \langle \{shirt_a,\, jeans_a\}, \{Above(shirt_a,\, jeans_a), Aligned(shirt_a,\, jeans_a)\} \rangle ,$$
$$G \rightarrow \langle \{sweater_a,\, jeans_a\}, \{Above(sweater_a,\, jeans_a), Aligned(sweater_a,\, jeans_a)\} \rangle ,$$

Note that all the spatial relations were considered in the same frame.

The grammar used in this work was built based on correct dressing examples. Thus, building the grammar requires specifying only the predicates, which takes a few minutes for each example; based on this information, then the grammar is completed automatically, without further intervention. Also it is easy to extend the system by adding predicates for other correct dressing examples, which can simply be appended to the existing grammar.

4.4 Step 4: Parsing a Dressing Image Sequence

The goal of parsing is to detect and explain failures in a dressing activity or label a sequence as correct dressing. There are three main types of failures that we consider, namely:

1. Temporal failures (wrong dressing order).
2. Spatial failures (putting on clothes partially).
3. Relational failures (putting on clothes backwards).

It should be noted that wrong dressing order is related to temporal relationships only. Putting clothes partially is a combination of temporal and spatial relationships. Putting on clothes backwards is a computer vision challenge. The first two failures are addressed by the grammar using the rule-based inference engine. The backwards problem is addressed by the lexicon of the grammar when the object backwards is detected in the image. Algorithm 1 shows the detection process of the three types of errors (temporal, spatial and relational error).

Algorithm 1: Error detection (parsing) with the proposed TR-grammar

Data: The rules $\mathcal{R}$ of a sequence example, the grammar $\mathcal{G}$ model
Result: A list of errors $\mathcal{E}_\mathcal{G}$ and its type.
foreach $r_i \in \mathcal{R}$ **do**
  foreach $s_i \in \mathcal{G}(S)$ **do**
    if $isNext(r_i)$ **then**
      if $Next(a,b) \in r_i and Next(b,a) \in s_i$ **then**
        addErr($\mathcal{E}_\mathcal{G}, r_i, Temp_{err}$);
      if $Next(a,b) \in s_i$ *and* $Next(a,a) \in r_i$ **then**
        addErr($\mathcal{E}_\mathcal{G}, r_i, Spat_{err}$);
    else
      if $r(a,b) \in r_i$ *and* $Next(a,b) \in s_i$ **then**
        addErr($\mathcal{E}_\mathcal{G}, r_i, Spat_{err}$);
      if $(a,b) \in r_i$ *and* $isBackwards(a) \in s_i$ *or* $isBackwards(b) \in s_i$ **then**
        addErr($\mathcal{E}_\mathcal{G}, r_i, Relat_{err}$);

Wrong dressing order failure The grammar detects wrong order of dressing by parsing temporal relationships only. If the example has a temporal relationship and this relationship does not appear in the learned grammar, the failure will be recognised by the parsing algorithm. In other words, we do not wear garments in the opposite way. For example, our parsing algorithm includes the following rule:

$$Next(shirt_a, jacket_{a+1}) \rightarrow \neg Next(jacket_a, shirt_{a+1}), \tag{1}$$

where a is a frame where the parser is operating. In Fig. 8, the error can be detected when the rule is violated. If the written rule $Next(shirt_a, jacket_{a+1})$ appears in the grammar, it implies that the opposite rule $Next(jacket_a, shirt_{a+1})$ should not appear in the example. If the rule exists in the example, the wrong dressing order failure will be detected.

Putting on clothes partially Our model detects failures in partial garment using temporal and spatial information. Firstly, we evaluate if the garment is placed in the correct position by using arity-one predicates. A failure of the wrong part of the body is explained using predicates declaring the spatial position of each garment:

$$G \rightarrow \langle \{shirt_a,\, jeans_a\}, \{Above(shirt_a,\, jeans_a), isUpper(shirt_a),\, isLower(jeans_a)\}\rangle, \tag{2}$$

The arity-one predicates solve the wrong part of the body problem: jeans cannot be in the upper position and shirts cannot be in the lower position. This rule indicates a strong restriction: the failure is detected when the arity-one predicates does not appear in a sequence. Afterwards, partial garment failures are addressed with the following rules:

$$Next(garmx_a, garmy_{a+1}) \rightarrow \neg Next(garmx_a, garmx_{a+1}) \tag{3}$$
$$Next(garmx_a, garmy_{a+1}) \rightarrow \neg Left(garmx_{a+1}, garmy_{a+1}), \tag{4}$$

In general, spatial relationships where the previous garment still appears on the second frame will be explained as a partial garment failure ($Left$, $Above$, $Aligned$, etc).

Putting on clothes backwards This problem was addressed by learning the texture of the garments in the backwards position, i.e. one garment was learned twice: once in its normal position and another in backwards position. Unfortunately, many garments have exactly the same texture on both sides. This is a difficult problem for current computer vision techniques; the model does not often obtain a correct classification of the garment, making it difficult to recognise this kind of failure. In terms of the lexicon, this requires to include a backwards garment detector. For example, if we have the $Shirt$ detector, we also should have the "$backwardsShirt$" detector. When backwardsShirt is detected, the failure is immediately detected as well. It should be noted that this kind of failure can be detected without the TR-grammar structure (only with the lexicon).

5 Results and Discussion

As far as we are aware, there has not been other work in automatically detecting dressing failures, therefore we present and compare our results with our previous work (Matic et al 2012), where we additionally used garments with RFID tags. As we show below, our results are comparable with our previous work, even though here we rely solely on the computer vision-based system, without using RFID data.

Table 1 The table summarizes the results as a confusion matrix between correct dressing and the different type of failures. Backwards was the most difficult case because many garments have the same texture and colour in the backwards position.

Event Type	Correct Dressing	Temporal failure	Spatial failure	Relational failure
Correct Dressing	80%	4%	16%	0%
Temporal failure	0%	80%	20%	0%
Spatial failure	0%	40%	40%	20%
Relational Failure	28.5%	14.3%	14.3%	42.9%

5.1 Dataset structure and test protocol

We used the same dataset as in Matic et al (2012) while excluding RFID information. The evaluation was carried out in terms of accuracy: a sequence of garments can be classified in four ways: correct dressing activity, temporal failure, spatial failure, and relational failure. Only one class is assigned to each sequence. The dataset consists of 47 sequences and each sequence has more than two images. Dimensions of each image are 1602 x 2848 pixels. This dataset has 25 correct examples and 22 examples of dressing failures: 10 temporal failures, 5 spatial failures and 7 relational failures. Evaluation was performed as follows:

- A sequence of images of arbitrary length, one for each dressing activity, is analysed (only image information is provided to the model).
- The grammar parses each image sequence and outputs an evaluation.
- Model evaluation provides either an explanation of the type of failure or labels the sequence as correct dressing activity.

5.2 Results

An overview of the results is presented in Table 1. Each row of the table shows the accuracy of our method for each class (the sum for each row is 100%). In the first row, an accuracy of 80% was achieved. In other words, 80% of the correct dressing activities were correctly parsed by the grammar. The rest were erroneously explained by the grammar as temporal or spatial failure. Majority of errors were caused by the imprecision of the vision system; relational failures were most challenging to recognise (42.9% of accuracy only). As it can be seen, there are a number of misclassifications in our model that are explained below.

Temporal failures There were several misclassified cases of temporal errors; that is, wrong order of garments. This was primarily because the second garment at times did not cover completely the previous garment, for example long sleeves in the first garment (such as the jumper) were not fully covered by the short sleeves in the second garment (such as t-shirt) suggesting partial dressing and consequently resulting in classifier confusion. In another case, a subject failed to wear a shirt in wrong order because the previous garment was a bulky jacket. The grammar processed the jacket in the last frame and suggested spatial failure.

Spatial failures As we have indicated in the introduction, we can less reliably detect spatial and relational failures. This is because several examples were not a clear-cut type of failure; that is, more than one type of failure could be observed.

In few instances where the garment has been put on partially (spatial failure) our system classified the dressing failure as backwards failure (relational failure), resulting in false positives. Other instances

Table 2 Precision and recall results when detecting dressing failures only, without considering type of failure. Precision shows that 9 out of 10 dressing failures can be detected reliably.

Precision & Recall in failures	
Precision	91%
Recall	77%
Accuracy	83%

Table 3 Confusion matrix with two classes. One class is for correct dressing and the other includes the three types of failures.

	correct dressing	failure
correct dressing	19	6
failure	2	20

involved temporal errors in partial garments; for example, in two cases the subject attempted to put on the garment but failed, resulting in partial garment failure. However, in addition to partial garment failure, a temporal failure also occurred; that is, the subject not only put on the garment partially, but also in the wrong order as it can be seen in Fig. 5 where the subject attempts to put on a polo shirt after a sweater. As such the TR-visual grammar identified it as temporal failure, but partial dressing failure also occurred.

Relational failures Backwards garment was considered (and learned) as another garment, since the classifier learned the texture of the garment put on backwards to distinguish it from correct dressing. However, there were several garments that had exactly the same texture on the inside as on the outside (for example shirts and t-shirts). Therefore, these cases could not be detected and explained well by the grammar. Clearly, this is a challenging issue to address, even for a human, where we typically look for seams of the garment, which was not possible to detect since seams in our images were less than a pixel small. Without the ability to detect texture changes, the results of backward failure are classified as correct dressing (28.5% accuracy in our experiments). In other cases, backwards failures are classified as partial garment or wrong order dressing. In both cases these failures were due to the difficulty of garment detection using only computer vision (which incidentally was one of the motivations of using RFID in our previous work). In particular, the second garment was a sweatshirt and was classified as t-shirt, giving rise to a partial garment error if the previous garment was of the same type, or temporal error if the t-shirt appeared before.

5.3 Detecting dressing failures only

Providing dressing failures only, without considering the type of failure, may be an important aspect in understanding the progression of a specific disease or improvement in patients' state through measuring number of dressing failures. In this respect, we are interested in a precision metric, measuring predictive value of dressing failures. Using our method, we achieve precision of 91% meaning that nine out of ten dressing failures can be detected as shown in Table 2, while sensitivity (recall) is 77% as shown in Table 2, along with the confusion matrix for each case in Table 3.

5.4 Efficiency

Once the model has been built, parsing an image sequence, including garment classification, takes few milliseconds (in a standard laptop with Intel(R) Core (TM) i7-6600U CPU @ 2.60GHz and 8.00 GB RAM). Thus, the proposed approach could be used to provide real-time on the dressing activity, while it could also be incorporated in other real-time applications such as passive monitoring scenarios.

Table 4 The table summarizes a NaïveBayes and SVM benchmark between correct dressing and the different type of failures. As one can expect, the ability to discover temporal and spatial failures is reduced. The first value corresponds to Naïve Bayes and the second one correspond to SVM with linear kernel

Event Type	Correct Dressing	Temporal failure	Spatial failure	Relational failure
Correct Dressing	68%/84%	4%/4%	28%/0%	0%/12%
Temporal failure	10%/30%	60%/50%	20%/10%	10%/10%
Spatial failure	60%/60%	40%/40%	0%/0%	0%/0%
Relational Failure	28.5%/100%	0%/0%	14.3%/0%	57.14%/0%

5.5 Comparison with Naïve Bayes and Support Vector Machine Classifiers

We implemented an alternative method for detection of failures in dressing activities based on standard classification techniques; that is by only considering the lexicon of the model, where the relational information (which in our model is described by the grammar) was omitted. A Naïve Bayes classifier and a Support Vector Machine (SVM) classifier were used to perform the classification task; the information of the lexicon was passed in the form of attributes. If a word of the lexicon appears in a frame it is added as a feature (the value is set to one when the garment was detected, and zero otherwise). The same test sequences were considered: 36 attributes were used for each sequence, 12 garments which can appear three times (in three frames). With this approach it is not possible to explain the failures, since the model can only detect dressing failures, but the model is unable to explain the error (i. e. what spatial or temporal relationship failed in what frame or frames). As one can expect, the model loses the knowledge about temporal information also. For the SVM classifier, a linear kernel was used, as other kernels had lower performance (RBF and Polykernel were tested). Table 4 summarises the results using tenfold cross validation.

These results show the benefits of including a visual grammar as it reduces the noise that is intrinsically present in the lexicon. The SVM classifier performed better for correct dressing; however, spatial, temporal and relational failures were not addressed well by this classification method. If we compare with the results of using TR visual grammar in Table 1, in general the performance is lower with both classifiers; additionally, neither can explain the failures.

5.6 Comparison with an SR-grammar

To highlight the advantages of the proposed temporal-relational (TR) grammar, in this section we provide a comparison with an SR-grammar for a synthetic KB example. The example is a relational description of waving two hands (to say hello with two hands). The actions that should be performed are summarized in table 5. The meaning of this knowledge base is: i) you should raise your hands together first, ii) you should wave your two hands and, iii) you should lower your two hands. These three actions are explained with relational structures. For each grammar, the description is written as follows:

Table 5 Syntethic knowledge-base to describe a gesture (waving with two hands).

Spatial Relationships		
$raise(Lhand_1)$	$wave(Lhand_2)$	$lower(Lhand_3)$
$raise(Rhand_1)$	$wave(Rhand_2)$	$lower(Rhand_3)$
$left(Lhand_2, Rhand_2)$		
Temporal Relationships		
$next(Lhand_1, Lhand_2)$	$next(Lhand_2, Lhand_3)$	
$next(Rhand_1, Rhand_2)$	$next(Rhand_2, Rhand_3)$	

– **TR-grammar**. According to our formalism, one way to write the grammar is:

$$W \rightarrow \langle \{Lhand_1,\, Rhand_1\}, \{raise(Lhand_1),\, raise(Rhand_1)\}\rangle \quad (5)$$
$$W \rightarrow \langle \{Lhand_2,\, Rhand_2\}, \{left(Lhand_2,\, Rhand_2), wave(Lhand_2),\, wave(Rhand_2)\}\rangle \quad (6)$$
$$W \rightarrow \langle \{Lhand_3,\, Rhand_3\}, \{lower(Lhand_3),\, lower(Rhand_3)\}\rangle \quad (7)$$
$$W \rightarrow \langle \{Lhand_1,\, Lhand_2\}, \{next(Lhand_1,\, Lhand_2)\}\rangle \quad (8)$$
$$W \rightarrow \langle \{Lhand_2,\, Lhand_3\}, \{next(Lhand_2,\, Lhand_3)\}\rangle \quad (9)$$
$$W \rightarrow \langle \{Rhand_1,\, Rhand_2\}, \{next(Rhand_1,\, Rhand_2)\}\rangle \quad (10)$$
$$W \rightarrow \langle \{Rhand_2,\, Rhand_3\}, \{next(Rhand_2,\, Rhand_3)\}\rangle, \quad (11)$$

where the subscripts refer the frame number where the object belongs to. Each connection in time is explicitly described for each production rule.

– **SR-grammar**. The SR-formalism provides a more cryptic writing:

$$WA^0 \rightarrow \left\langle \left\{Lhand^2,\, Rhand^2\right\}, \left\{raise(Lhand^2),\, raise(Rhand^2)\right\}\right\rangle \quad (12)$$
$$WB^0 \rightarrow \left\langle \left\{Lhand^2,\, Rhand^2\right\}, \left\{left(Lhand^2,\, Rhand^2), wave(Lhand^2),\, wave(Rhand^2)\right\}\right\rangle \quad (13)$$
$$WC^0 \rightarrow \left\langle \left\{Lhand^2,\, Rhand^2\right\}, \left\{lower(Lhand^2),\, lower(Rhand^2)\right\}\right\rangle \quad (14)$$
$$WT^0 \rightarrow \left\langle \left\{WA^2,\, WB^2\right\}, \left\{next(WA^2,\, WB^2)\right\}\right\rangle \quad (15)$$
$$WTT^0 \rightarrow \left\langle \left\{WT_2,\, WC^2\right\}, \left\{next(WT^2,\, WC^2)\right\}\right\rangle, \quad (16)$$

where the initial symbol is WTT. Since we do not have information about the frames, we have to put the information in non-terminal elements. This makes this description more difficult to build and interpret.

This example illustrates the advantages of the proposed TR grammar; it facilitates building and interpreting a description that includes spatial and temporal relations. This could help to reduce errors when defining a grammar for practical applications.

5.7 Discussion and comparison with previous work

In our previous work RFID tags were used to tag each garment and RFID antennas were mounted inside the dressing area to obtain additional spatial and temporal information. In this work, we use visual information only, since the garments are learned by visual classifiers and the inference is performed by the TR-visual grammar. These results are comparable to the previous results (where we used RFID tags) at 80% versus 83.9% respectively for correct dressing. In terms of dressing failures, wrong order has the same detection performance (80%), although TR grammars recognised misclassified examples as partial garment failure, whereas RFID determined the misclassified sequences as unrecognised. In partial dressing the previous work performed better, however as we stated in the previous section, a number of errors in the grammar occurred due to confusion with the temporal failures and false positives in the vision system with backwards garments. Detecting backwards failure had better performance when using RFID, 83.3% vs. 42.9%, which was expected since detection of garment backwards is much easier with RFID than using vision only: RFID tags were detected in the opposite positions when the garment was put on backwards. As such, the errors in our grammar model were due to garments with similar texture.

5.7.1 Directions for improvement

The errors in our system are mainly due to failures in the vision-based garment detectors. These detectors, which are not the main focus of this work, could be improved in several ways, such as: (i) incorporate local features, such as SIFT, in addition to the global features used here; (ii) test other classifiers, such as Random Forests (Breiman 2001); (iii) detect garments' brand labels, which could improve the classification of backwards clothes; (iv) incorporate deep learning techniques for garment representation and recognition. Finally, if we focus on detecting whether a dressing failure has occurred, without being concerned with the type of failure, then we achieve precision of 91% which means that our system can detect failures even when sometimes these failures are not well explained. The source of errors stems from two principal aspects, namely discriminating between temporal and spatial failures, and detecting garments backwards.

6 Conclusion and Future Work

In this paper we describe a novel grammar-based method to recognise dressing failures and their type. The grammar has two main contributions: (i) the expressive power of knowledge representation, and (ii) the combination of spatial and temporal relations. This expressive power is used to recognise failures in dressing activities and explain types of dressing failures. The experimental evaluation shows that the proposed vision-based method can distinguish between a failure and correct dressing with 91% precision. This is the first work to investigate the automatic detection of dressing failures relying only on visual information.

There are several avenues to pursue in the future work. One is the improvement of garment detection. We based our method on vision only, as it is one of the least expensive and most practical methods. Applying our model in real-time video is a future avenue that may improve the results, because there is additional information that can be used to improve garment recognition.

Another interesting challenge would be to deduce a grammar from several examples. Finally, we are also interested in exploring how to handle uncertainty in domains where the knowledge representation is not always true or false: a rule-based algorithm is not appropriate when there is uncertainty in the recognition activity (vision-based detectors are not perfect). As such, Probabilistic Graphical Models or Statistical Relational Models might be considered.

References

Afsar P, Cortez P, Santos H (2015) Automatic visual detection of human behavior: a review from 2000 to 2014. Expert Systems with Applications

Allen JF (1983) Maintaining knowledge about temporal intervals. Commun ACM 26(11):832–843, DOI 10.1145/182.358434, URL http://doi.acm.org/10.1145/182.358434

Bahle G, Gruenerbl A, Lukowicz P, Bignotti E, Zeni M, Giunchiglia F (2014) Recognizing hospital care activities with a coat pocket worn smartphone. In: Mobile Computing, Applications and Services (MobiCASE), 2014 6th International Conference on, IEEE, pp 175–181

Bay H, Ess A, Tuytelaars T, Gool LJV (2008) Speeded-up robust features (SURF). Computer Vision and Image Understanding 110(3):346–359, DOI 10.1016/j.cviu.2007.09.014, URL http://dx.doi.org/10.1016/j.cviu.2007.09.014

Breiman L (2001) Random forests. Machine Learning 45(1):5–32, DOI 10.1023/A:1010933404324, URL http://dx.doi.org/10.1023/A:1010933404324

Brown C, Moore WP, Hemman D, Yunek A (1996) Influence of instrumental activities of daily living assessment method on judgments of independence. American Journal of Occupational Therapy 50(3):202–206

Chen C, Zhang D, Sun L, Hariz M, Yuan Y (2012) Does location help daily activity recognition? In: Impact Analysis of Solutions for Chronic Disease Prevention and Management, Springer, pp 83–90

Chernbumroong S, Cang S, Atkins A, Yu H (2013) Elderly activities recognition and classification for applications in assisted living. Expert Systems with Applications 40(5):1662–1674

Chomsky N (2002) Syntactic structures. A Mouton classic, Mouton de Gruyter, URL http://books.google.com/books?id=SNeHkMXHcd8C

Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20(3):273–297, DOI 10.1007/BF00994018, URL http://dx.doi.org/10.1007/BF00994018

Costagliola G, Deufemia V, Ferrucci F, Gravino C (2002) Using extended positional grammars to develop visual modeling languages. In: Proceedings of the 14th international conference on Software engineering and knowledge engineering, SEKE 2002, Ischia, Italy, July 15-19, 2002, ACM, pp 201–208, DOI 10.1145/568760.568795, URL http://doi.acm.org/10.1145/568760.568795

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA, IEEE Computer Society, pp 886–893, DOI 10.1109/CVPR.2005.177, URL http://dx.doi.org/10.1109/CVPR.2005.177

Edmans J, Lincoln N (1990) The relation between perceptual deficits after stroke and independence in activities of daily living. The British Journal of Occupational Therapy 53(4):139–142

Edmans J, Towle D, Lincoln NB (1991) The recovery of perceptual problems after stroke and the impact on daily life. Clinical rehabilitation 5(4):301–309

Ferrucci F, Pacini G, Satta G, Sessa MI, Tortora G, Tucci M, Vitiello G (1996) Symbol-relation grammars: a formalism for graphical languages. Inf Comput 131(1):1–46, DOI http://dx.doi.org/10.1006/inco.1996.0090

Feyereisen P (1999) Disorders of everyday actions in subjects suffering from senile dementia of alzheimerś type: An analysis of dressing performance. Neuropsychological Rehabilitation 9(2):169–188

Foncubierta-Rodríguez A, Müller H, Depeursinge A (2017) From visual words to a visual grammar: using language modelling for image classification. CoRR abs/1703.05571, URL http://arxiv.org/abs/1703.05571, 1703.05571

Friedman A, Ron S (2017) Unlocking the power of visual grammar theory: analyzing social media political advertising messages in the 2016âĂŞus election. Journal of Visual Literacy 36(2):90–103, DOI 10.1080/1051144X.2017.1379758, URL https://doi.org/10.1080/1051144X.2017.1379758, https://doi.org/10.1080/1051144X.2017.1379758

Girshick RB, Felzenszwalb PF, McAllester DA (2011) Object detection with grammar models. In: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira FCN, Weinberger KQ (eds) NIPS, pp 442–450

Gottfried B (2015) The systematic design of visual languages applied to logical reasoning. J Vis Lang Comput 28:212–225, DOI 10.1016/j.jvlc.2015.02.001, URL http://dx.doi.org/10.1016/j.jvlc.2015.02.001

Haralick RM, Shanmugam KS, Dinstein I (1973) Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics 3(6):610–621, DOI 10.1109/TSMC.1973.4309314, URL http://dx.doi.org/10.1109/TSMC.1973.4309314

Kong J, Zhang K, Zeng X (2006) Spatial graph grammars for graphical user interfaces. ACM Trans Comput-Hum Interact 13(2):268–307, DOI 10.1145/1165734.1165739, URL http://doi.acm.org/10.1145/1165734.1165739

Lakin F (1987) Visual grammars for visual languages. In: Forbus KD, Shrobe HE (eds) Proceedings of the 6th National Conference on Artificial Intelligence. Seattle, WA, July 1987., Morgan Kaufmann, pp 683–688, URL http://www.aaai.org/Library/AAAI/1987/aaai87-122.php

Leborg C (2006) Visual grammar. Princeton Architectural Press

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2):91–110

Maio CD, Fenza G, Gallo M, Loia V, Parente M (2017a) Time-aware adaptive tweets ranking through deep learning. Future Generation Computer Systems DOI https://doi.org/10.1016/j.future.2017.07.039, URL http://www.sciencedirect.com/science/article/pii/S0167739X17308087

Maio CD, Fenza G, Loia V, Orciuoli F (2017b) Unfolding social content evolution along time and semantics. Future Generation Comp Syst 66:146–159, DOI 10.1016/j.future.2016.05.039, URL https://doi.org/10.1016/j.future.2016.05.039

Marriott K, Meyer B (1996) Towards a hierarchy of visual languages. In: Proceedings of the 1996 IEEE Symposium on Visual Languages, Boulder, Colorado, USA, September 3-6, 1996, IEEE Computer Society, pp 196–203, DOI 10.1109/VL.1996.545288, URL http://dx.doi.org/10.1109/VL.1996.545288

Matic A, Mehta P, Rehg JM, Osmani V, Mayora O (2010) Aid-me: Automatic identification of dressing failures through monitoring of patients and activity evaluation. In: Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2010 4th International Conference on, IEEE, pp 1–8

Matic A, Mehta P, Rehg JM, Osmani V, Mayora O (2012) Monitoring dressing activity failures through rfid and video. Journal of Methods of Information in Medicine 51:45–54

Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630, DOI 10.1109/TPAMI.2005.188, URL http://doi.ieeecomputersociety.org/10.1109/TPAMI.2005.188

Mjolsness E (1991) Visual grammars and their neural nets. In: Moody JE, Hanson SJ, Lippmann R (eds) Advances in Neural Information Processing Systems 4, [NIPS Conference, Denver, Colorado, USA, December 2-5, 1991], Morgan Kaufmann, pp 428–435, URL http://papers.nips.cc/paper/499-visual-grammars-and-their-neural-nets

Namazi KH, Johnson BD (1992) Dressing independently: A closet modification model for alzheimerś disease patients. American Journal of Alzheimerś Disease and Other Dementias 7(1):22–28

Ojala T, Pietikäinen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distributions. Pattern Recognition 29(1):51–59, DOI 10.1016/0031-3203(95)00067-4, URL http://dx.doi.org/10.1016/0031-3203(95)00067-4

Osmani V (2015) Smartphones in mental health: Detecting depressive and manic episodes. IEEE Pervasive Computing 14(3):10–13

Qi S, Huang S, Wei P, Zhu S (2017) Predicting human activities using stochastic grammar. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, pp 1173–1181, DOI 10.1109/ICCV.2017.132, URL https://doi.org/10.1109/ICCV.2017.132

Rublee E, Rabaud V, Konolige K, Bradski GR (2011) ORB: an efficient alternative to SIFT or SURF. In: Metaxas DN, Quan L, Sanfeliu A, Gool LJV (eds) IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011, IEEE, pp 2564–2571, DOI 10.1109/ICCV.2011.6126544, URL http://dx.doi.org/10.1109/ICCV.2011.6126544

Sunderland A, Walker CM, Walker MF (2006) Action errors and dressing disability after stroke: an ecological approach to neuropsychological assessment and intervention. Neuropsychological rehabilitation 16(6):666–683

Vedaldi A, Fulkerson B (2010) Vlfeat: An open and portable library of computer vision algorithms. In: Proceedings of the 18th ACM International Conference on Multimedia, ACM, New York, NY, USA, MM '10, pp 1469–1472, DOI 10.1145/1873951.1874249, URL http://doi.acm.org/10.1145/1873951.1874249

Walker C, Walker M (2001) Dressing ability after stroke: a review of the literature. The British Journal of Occupational Therapy 64(9):449–454

Walker CM, Walker MF, Sunderland A (2003) Dressing after a stroke: a survey of current occupational therapy practice. The British Journal of Occupational Therapy 66(6):263–268

Walker M, Lincoln N (1991) Factors influencing dressing performance after stroke. Journal of Neurology, Neurosurgery & Psychiatry 54(8):699–701

Walker M, Drummond A, Lincoln N (1996) Evaluation of dressing practice for stroke patients after discharge from hospital: a crossover design study. Clinical Rehabilitation 10(1):23–31

Wittenburg K, Weitzman L (1996) Relational grammars: Theory and practice in a visual language interface for process modeling. In: In Workshop on Theory of Visual Languages, Springer Verlag, pp 27–29

Wu YN, Si Z, Gong H, Zhu SC (2010) Learning active basis model for object detection and recognition. International Journal of Computer Vision 90(2):198–235

Zhu L, Chen Y, Yuille AL (2009) Unsupervised learning of probabilistic grammar-markov models for object categories. IEEE Trans Pattern Anal Mach Intell 31(1):114–128, DOI 10.1109/TPAMI.2008.67, URL http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.67

Zhu SC, Mumford D (2006) A stochastic grammar of images. Foundations and Trends in Computer Graphics and Vision 2(4):259–362, URL http://dx.doi.org/10.1561/0600000018