A Hierarchical Human Activity Recognition Framework Based on Automated Reasoning

Shuwei Chen, Jun Liu, Hui Wang
School of Computing and Mathematics
University of Ulster at Jordanstown
Newtownabbey, BT37 0QB, UK
chen-s1@email.ulster.ac.uk, j.liu@ulster.ac.uk,
h.wang@ulster.ac.uk

Juan Carlos Augusto
Department of Computer Science
Middlesex University
London, UK
j.augusto@mdx.ac.uk

***Abstract*—Conventional human activity recognition approaches are mainly based on machine learning methods, which are not working well for composite activity recognition due to the complexity and uncertainty of real scenarios. We propose in this paper an automated reasoning based hierarchical framework for human activity recognition. This approach constructs a hierarchical structure for representing the composite activity by a composition of lower-level actions and gestures according to its semantic meaning. This hierarchical structure is then transformed into logical formulas and rules, based on which the resolution based automated reasoning is applied to recognize the composite activity given the recognized lower-level actions by machine learning methods.**

Keywords—human activity recognition; hierarchical approach; resolution priciple; automated reasoning

I. INTRODUCTION

Video based human activity recognition is attracting more and more attention now due to the vast needs from a variety of application systems, such as video surveillance systems, healthcare systems [1]. A generic activity recognition system can be viewed as proceeding from a sequence of images to a higher level interpretation in a series of steps [2]. The major steps involved are the following:

1) Input video or sequence of images;
2) Extraction of concise low-level features;
3) Mid-level action descriptions from low-level features;
4) High-level semantic interpretations from primitive actions.

There are generally two types of approaches for activity recognition: single-layered approaches and hierarchical approaches [1]. Single-layered approaches represent and recognize human activities directly from video data by considering an activity as a particular class of image sequences. Due to their nature, single-layered approaches are suitable for the recognition of gestures and actions with sequential characteristics, such as walking, hand waving, and running. On the other hand, hierarchical approaches represent high-level human activities by describing them in terms of other simpler activities, generally called sub-events, with the assumption that the simpler activities can be relatively easily recognized first. A hierarchical recognition system is composed of multiple layers, making them suitable for the analysis of complex activities, such as fighting, and people meet. The complex activities are generally classified into three types: composite actions, interactions and group activities [1].

The current computer vision approaches, such as hidden Markov models (HMM), dynamic Bayesian network (DBN), Support Vector Machines (SVM) etc. [1, 3, 4], work well for the extraction and recognition of features, gestures and simpler actions. However, they are limited in the case of recognizing high-level activities due to the uncertainty and complex nature of human movement. For these situations, it is hard to define a general motion sequence to allow the use of a general sequence matching approach [5]. To overcome this limitation, prior knowledge should be considered for human activity recognition. As a result, hierarchical approaches, which are suitable for a semantic-level analysis of interactions between humans and/or objects as well as complex group activities, have been studied for complex human activity recognition [1, 3, 6].

Despite the fact that extensive efforts have been devoted recent years, bridging the semantic gap between low level data and high level human understanding is still a challenge [7]. Many available hierarchical approaches have limited flexibility, or have difficulty with the computational complexity of composite activity recognition tasks. The lack of effective reasoning mechanisms from low level data to high level semantic understanding limits the ability for recognizing complex activities in real world applications.

Following the similar idea of hierarchical task network (HTN) [8, 9], which decomposes the tasks to be performed into simpler subtasks until primitive tasks or actions that can be directly executed can be reached, this paper proposes a framework which represents the composite activity under consideration by a hierarchical structure and recognizes composite activity using resolution based automated reasoning. This hierarchical structure generally consists of three layers: low-level features and gestures, mid-level actions, and high-level activity, which is constructed based on the knowledge of the semantic meaning of the considered activity. This hierarchical structure is then transformed into logical formulas and rules. During low-level (or atomic) recognition, image

978-1-4799-0652-9/13 $31.00 © 2013 IEEE
DOI 10.1109/SMC.2013.596

sequences are processed by conventional computer vision methods to identify human actions and gestures. Then, a resolution based automated reasoning method is applied to recognize the composite activity based on the transformed logical formulas and the recognized lower-level actions.

The rest of this paper is structured as follows. The proposed activity recognition framework and the hierarchical structure for representing composite activity are presented in Section 2. This hierarchical structure is then transformed into logical formulas and rules in Section 3, based on which the resolution based automated reasoning is applied to recognize the composite activity given the recognized lower-level actions by conventional machine learning methods. An illustrative example is given in Section 4 to show the proposed approach, and conclusions and discussions are drawn in Section 5.

II. Hierarchical Representation Structure for Activity Recognition

The hierarchical framework of the composite activity recognition approach consists of generally three layers, and can be illustrated as Fig. 1. This hierarchical framework is constructed according to the semantic understanding of the activities, which enables the users apply prior knowledge to the activity recognition process. It decomposes the high-level composite activity to mid-level actions, and then to low-level gestures and features by applying the human understanding of the activity. The composite activity recognition is then achieved through a bottom-up process. Firstly, object and person detections, along with low-level feature extraction and atomic action recognition are realized via machine learning methods. Subsequently, the hierarchical structure is transformed into logical formula representation and the automated reasoning method is applied to inference from the detected lower-level actions and gestures to the high-level composite activity. The paradigm of hierarchical representation not only makes the recognition process computationally tractable and conceptually understandable, but also makes the recognition process more effective by applying automated reasoning mechanism.

In order to illustrate the hierarchical representation structure for modeling the considered activity, we take embrace interaction between two persons as an example whose representation structure is shown as Fig. 2. For simplicity, the input video level and the image features are not shown here.

Low-level feature and gesture recognition are achieved through HMM classification using optical flow features, which have been proposed and evaluated previously [10]. The inputs for low-level feature and gesture recognition are regions of interest corresponding to tracked persons within a video scene, and may be generating using methods such as foreground (GMM) modelling. For each region of interest we generate optical flow features capturing motion orientation, magnitude, and relative location (as described previously [10]), resulting in 154*1 features per person detection per frame. Each sequence of feature vectors are used as inputs for classification, and each observation sequence is labelled based on the HMM log-likelihood scores for each possible atomic gesture and action.

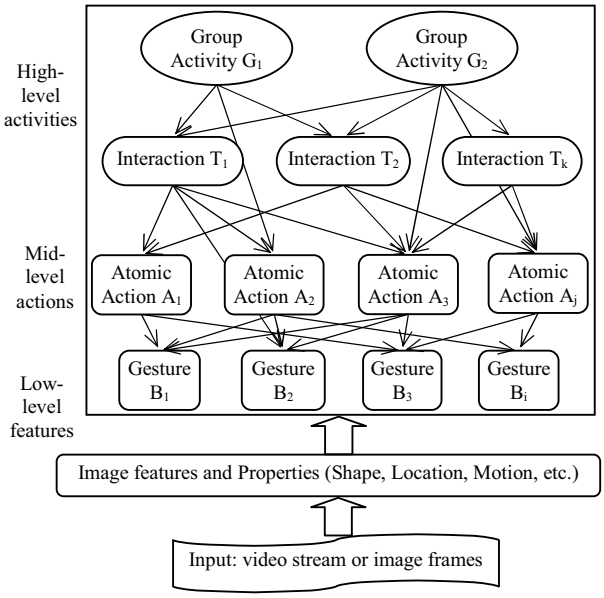

Figure 1. The hierarchical framework for activity recognition

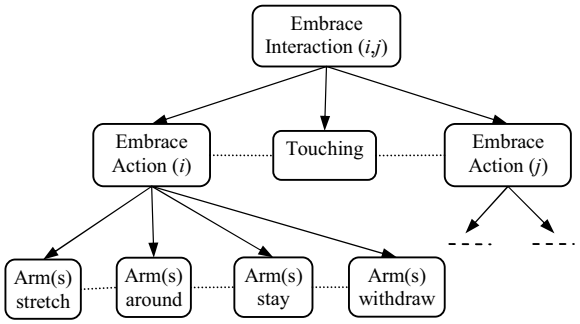

Figure 2. The hierarchical structure of embrace interaction

As discussed above, the low-level features and gestures, and some simpler actions can be well recognized by the existing machine learning methods, so we mainly focus on the recognition process from the low-level features and gestures to the mid-level actions and then to the high-level activity in the following. The most straightforward way for this process is matching, while we propose to use a more effective way, automated reasoning, to achieve this goal. Furthermore, the recognition result is more reliable due to the fact that the automated reasoning method has a strict logic foundation.

III. Automated Reasoning Based Activity Recognition

In this section, we propose to use *resolution based automated reasoning* for activity recognition problem which will provide an automated recognition process. This approach verifies whether the activity under consideration holds or not based on provided recognized primitive actions and the logical formulas and rules transformed from the hierarchical structure presented in Section 2. Some preliminary knowledge about resolution principle is provided as follow, and the readers may refer to [11] for more details.

A. Resolution Based Automated Reasoning

Resolution principle was first proposed by Robinson [12] for automated theorem proving, and further extensively studied in the context of finding natural and efficient proof systems to support a wide spectrum of computational tasks [11]. In essence, resolution-based automated theorem proving proceeds by constructing refutation proofs, i.e., proofs by contradiction. In other words, it transforms the proving of the validity of a theorem into validating the unsatisfiability of a logical formula variation from this theorem. Then a resolution algorithm is constructed to prove the unsatisfiability of this logical formula.

Definition 1 (Resolvent) For any two clauses (disjunction of literals) C_1 and C_2, if there is a literal L_1 in C_1 that is complementary to a literal L_2 in C_2, then delete L_1 and L_2 from C_1 and C_2 respectively, and construct the disjunction of the remaining clauses. The constructed clause is a resolvent of C_1 and C_2.

Theorem 1 Given the two clauses C_1 and C_2, a resolvent C of C_1 and C_2 is a logical consequence of C_1 and C_2. That is, if both C_1 and C_2 are true, then the resolvent C must be true. Conversely, if C is false (especially an empty set), then at least one of C_1 and C_2 is false. Given the fact that one of them, say C_1, is true, then we can conclude that another, C_2, is false.

Resolution Principle Given a set S of clauses, a (resolution) deduction of C from S is a finite sequence C_1, C_2, …, C_n of clauses such that either C is a clause in S or a resolvent of clauses preceding C; and C_n=C. A deduction of an empty set from S is called a refutation or a proof of S.

According to the resolution principle, proving the unsatisfiability of a logical formula is to construct a refutation of it by using a resolution algorithm. The formulas are basically sets of clauses each of which is a disjunction of literals, and the forms of literals are simple because they usually contain neither constants nor implication connectives. The resolution algorithm for implementing resolution principle usually simplifies judging if two literals being resolvent into judging if the two literals are the complementary pair.

B. Activity Recognition Based on Automated Reasoning

The hierarchical structure as illustrated in Fig. 2 provides an intuitive description of the activity under consideration based on its semantic meaning. In order to further proceed the recognition task, we need to use some symbols to represent the activity and its relations with its lower-level actions or gestures.

In this paper, we treat the activity (actually its negation) to be recognized as the high-level node, which is represented as the disjunctive normal forms of the mid-level action nodes, and the mid-level action nodes are consequently the disjunctive normal forms of the low-level gesture or feature nodes denoted as literals (atomic formulas or their negations). This process is semantically nature due to the fact that the high-level activity is essentially composed some lower level actions and gestures. In other words, although the hierarchical structure is constructed according to the semantic understanding of the activity, it naturally has a close relationship to logic representation. Take the embrace action illustrate in Fig. 2 as an example, denote the arm stretch gesture as q_1, arm around as q_2, arm stay as q_3, and arm withdraw as q_4, then the embrace action of one person is $f = q_1 \wedge q_2 \wedge q_3 \wedge q_4$, whose negation is a disjunctive normal form $\neg f = \neg q_1 \vee \neg q_2 \vee \neg q_3 \vee \neg q_4$.

The logical formula as shown above is actually the symbolic representation of the semantic hierarchical structure as illustrated in Fig. 2. It also reflects the relationship between the high-level activity and its lower-level gestures or actions. The resolution algorithm is then applied to inference automatically whether this high-level logical formula holds or not, based on the recognized lower-level gestures and actions which can be seen as the inputs of the resolution algorithm.

The steps of the automated reasoning based hierarchical activity recognition process are summarized as follows:

Step 1. Identify problem domain, including possible scenarios, possible activities, and so on.

Step 2. Construct the hierarchical representation structure, as illustrated in Fig. 2, of the considered activity according to the prior knowledge of its semantic meaning.

Step 3. Represent low-level gestures as literals, and the mid-level actions and high-level activities as clauses and disjunctive normal forms based on the hierarchical representation structure constructed in Step 2.

Step 4. Recognize gestures and lower-level actions based on existing computer vision methods, such as HMM and DBN.

Step 5. Given recognized gestures and the logical disjunctive normal forms, apply resolution based automated reasoning to verify whether the activity is true or not, which is made by proving whether its negation is false or not.

Next section gives an example to illustrate the proposed automated reasoning based hierarchical activity recognition approach.

IV. Illustrative Example

We take the embrace interaction as an example to illustrate the proposed activity recognition approach. The hierarchical structure has already been constructed as in Fig. 2, so we start from Step 3.

Step 3. Symbolic (logical formula) representation

Denote $F(x, y)$: embrace interaction of person x and person y, $f(x)$: embrace action of person x, $h(x, y)$: person x is touching person y, $q_1(x)$: arm stretch of person x, $q_2(x)$: arm around of x, $q_3(x)$: arm stay of x, $q_4(x)$: arm withdraw of x.

According to the underlying semantic meaning of the hierarchical structure of the embrace interaction, there are three logical formulas (rules) involved:

$$(f(x) \wedge h(x, y)) \vee (f(y) \wedge h(y, x)) \rightarrow F(x, y), \quad (1)$$

$$q_1(x) \wedge q_2(x) \wedge q_3(x) \wedge q_4(x) \rightarrow f(x), \quad (2)$$

$$q_1(y) \wedge q_2(y) \wedge q_3(y) \wedge q_4(y) \rightarrow f(y). \quad (3)$$

The recognition task is to recognize the embrace interaction $F(x, y)$, which is transformed to validate the unsatisfiability of its negation $\neg F(x, y)$ according to the proposed method. Furthermore, the unsatisfiability of $\neg F(x, y)$ is equivalent to the unsatisfiability of $\neg((f(x) \wedge h(x, y)) \vee (f(y) \wedge h(y, x)))$ according to (1), which can be simplified into

$$(\neg f(x) \vee \neg h(x, y)) \wedge (\neg f(y) \vee \neg h(y, x)). \quad (4)$$

In order to verify the unsatisfiability of (4) using resolution based automated reasoning, it is divided into two clauses, which are denoted as: $C_1 = \neg f(x) \vee \neg h(x, y)$, and $C_2 = \neg f(y) \vee \neg h(y, x)$. By incorporating (2) and (3), we have

$$C_1 = \neg q_1(x) \vee \neg q_2(x) \vee \neg q_3(x) \vee \neg q_4(x) \vee \neg h(x, y), \quad (5)$$

$$C_2 = \neg q_1(y) \vee \neg q_2(y) \vee \neg q_3(y) \vee \neg q_4(y) \vee \neg h(y, x). \quad (6)$$

Hence, the recognition task for recognizing the embrace interaction $F(x, y)$ has been transformed into validating the unsatisfiability of the set S of clauses C_1 and C_2. Of course, new valid clauses, i.e., recognized gestures and simple actions, must be added to this set, otherwise we are not able to achieve the recognition task.

Step 4. Low-level gesture and action recognition

Based on the low-level feature and gesture recognition approach proposed in [10], which is based on HMM classification using optical flow features, we can recognize the gestures, and sometimes simple actions from the video (a sequence of image frames).

Suppose that we have detected that there are two persons, denoted as a and b, in the video, and the arm gestures $q_1(a)$, $q_2(a)$, $q_3(a)$, $q_4(a)$, $q_1(b)$, $q_3(b)$, $q_4(b)$, while the arm around gesture of person b, $q_2(b)$, is not detected due to some unknown reason. Furthermore, the touching actions, $h(a, b)$ and $h(b, a)$, have also been detected. These gestures and actions are then added as new clauses to the set of clauses S in Step 3.

Step 5. Automated reasoning based activity recognition

It can be seen from (5) and (6) that the clauses C_1 and C_2 are essentially the same, so we need only to keep one of them, say C_1, for the resolution based automated reasoning process, which is shown as follows.

1 $C_1 = \neg q_1(x) \vee \neg q_2(x) \vee \neg q_3(x) \vee \neg q_4(x) \vee \neg h(x, y)$

2 $q_1(a)$

3 $q_2(a)$

4 $q_3(a)$

5 $q_4(a)$

6 $q_1(b)$

7 $q_3(b)$

8 $q_4(b)$

9 $h(a, b)$

10 $h(b, a)$

11 $\neg q_2(a) \vee \neg q_3(a) \vee \neg q_4(a) \vee \neg h(a, y)$ (1+2)

12 $\neg q_3(a) \vee \neg q_4(a) \vee \neg h(a, y)$ (11+3)

13 $\neg q_4(a) \vee \neg h(a, y)$ (12+4)

14 $\neg h(a, y)$ (13+5)

15 $\neg q_2(b) \vee \neg q_3(b) \vee \neg q_4(b) \vee \neg h(b, y)$ (1+6)

16 $\neg q_2(b) \vee \neg q_4(b) \vee \neg h(b, y)$ (15+7)

17 $\neg q_2(b) \vee \neg h(b, y)$ (16+8)

18 $\neg q_2(b)$ (17+10)

19 $\square$ (empty set) (14+9)

Formulas (1)-(10) are the original clauses in the set S, and the following formulas are resolvents of some former clauses whose numbers are shown after the resolvents. When obtaining formula (11), the replacement of free variable x by constant a is called *substitution* in resolution based automated reasoning.

The deduction of the empty set in step 15 shows that we have obtained a refutation of the set S, and this means that we have validated the unsatisfiability of S, further the unsatisfiability of $\neg F(a, b)$. It means that $F(a, b)$ holds, i.e., there is embrace interaction between person a and person b in the video.

It can be seen that we can still achieve the recognition result although the arm around gesture of person b, $q_2(b)$, is not detected. Actually, the detected gestures of person b essentially make no sense during the reasoning process. The reason is that the logical relation between the embrace actions is disjunction as shown in (1) according to the human understanding of embrace interaction. This further shows that the proposed method is a knowledge-based approach.

Note that the above resolution process is done manually which is just to illustrate the proposed approach. In fact, there are many resolution based automated reasoning algorithms [11] that can be used to recognize more complex activities automatically and effectively.

V. Conclusions

This paper has presented a hierarchical framework for human activity recognition, which represents the composite activity by a hierarchical structure and recognizes the activity through resolution based automated reasoning. This approach constructs the hierarchical representation structure from the semantic point of view, and fulfills the recognition process through logic based automated reasoning way instead of straightforward matching. Therefore, it is helpful to bridge the semantic gap between low level data and high level human understanding, and provide a more effective and reliable way for complex activity recognition.

Although the proposed framework is mainly focus on video based activity recognition, it can be applied to sensor based activity recognition with some minor adjustments. The

proposed activity recognition approach is a general framework, and there is much work to be done. We will take uncertainty, temporal and spatial issues into consideration in future work. More comparison study with other existing methods and more experimental evaluations are required for future work.

ACKNOWLEDGMENT

This work has been partially supported by the VCRS scholarship from University of Ulster, the research project TIN2012-31263, and the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 285621, project titled SAVASA.

REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: a review," ACM Computing Surveys, vol. 43, no. 3, Article 16. 2011.

[2] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," IEEE Trans. Circuits Syst. Video Technol. Vol. 18, no. 11, pp.1473–1488, 2008.

[3] P. Natarajan and R. Nevatia, "Hierarchical multi-channel hidden semi Markov graphical models for activity recognition," Computer Vision and Image Understanding, in press, 2012, doi: http://dx.doi.org/10.1016/j.cviu.2012.08.011

[4] S. Christian, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," Proceedings of the International Conference on Pattern Recognition, vol. 3, pp. 32–36, 2004.

[5] S. Kim, H. S. Choi, K. M. Yi, J. Y. Choi, and S. G. Kong, "Intelligent visual surveillance - A survey," International Journal of Control, Automation, and Systems, vol. 8, no. 5, pp. 926–939, 2010.

[6] L. Wang, T. Gu, X. Tao, and J. Lu, "A hierarchical approach to real-time activity recognition in body sensor networks," Pervasive and Mobile Computing, vol. 8, pp. 115–130, 2012.

[7] L. Ballan and M. Bertini, "Video annotation and retrieval using ontologies and rule learning", IEEE Multimedia, vol. 17, no. 4, pp. 80–88, 2010.

[8] K. Erol, J. Hendler, and D. S. Nau, "HTN planning: complexity and expressivity," Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94), vol. 2, pp. 1123–1128, 1994.

[9] L. Macedo and A. Cardoso, "Case-based, decision-theoretic, HTN planning," Proceedings of Advances in Case-Based Reasoning, 7th European Conference, ECCBR 2004, pp. 257–271, 2004.

[10] S. Little and I. Jarglsaikhan, "SAVASA project at TRECVID 2012: Interactive Surveillance Event Detection," NIST TRECVID Workshop, 2012.

[11] G. J. Wang and H. J. Zhou, Introduction to Mathematical Logic and Resolution Principle, 2nd ed., Oxford: Alpha Science International Limited, 2009.

[12] J. A. Robinson, "A machine-oriented logic based on the resolution principle," Journal of the ACM, vol. 12, no. 1, pp. 23–41, 1965.