Human Action Recognition in Video by Fusion of Structural and Spatio-temporal Features

Ehsan Zare Borzeshi[1], Oscar Perez Concha[2], and Massimo Piccardi[1]

[1] School of Computing and Communications, Faculty of Engineering and IT, University of Technology, Sydney (UTS), Sydney, Australia
{ehsan.zareborzeshi, massimo.piccardi}@uts.edu.au
[2] Centre for Health Informatics, Australian Institute of Health Innovation, University of New South Wales, Sydney (UNSW), Australia
o.perezconcha@unsw.edu.au

Abstract. The problem of human action recognition has received increasing attention in recent years for its importance in many applications. Local representations and in particular STIP descriptors have gained increasing popularity for action recognition. Yet, the main limitation of those approaches is that they do not capture the spatial relationships in the subject performing the action. This paper proposes a novel method based on the fusion of global spatial relationships provided by graph embedding and the local spatio-temporal information of STIP descriptors. Experiments on an action recognition dataset reported in the paper show that recognition accuracy can be significantly improved by combining the structural information with the spatio-temporal features.

Keywords: Graph, Graph embedding, Human action recognition, STIP, Markov models.

1 Introduction and Related Work

Human action recognition has been the focus of much recent research for its increasing importance in applications such as video surveillance, human-computer interaction, multimedia and others. Recognising human actions is a challenging task, especially when the background is not fixed or known and the lighting conditions are changeable. Local representations and in particular appearance descriptors centred around spatio-temporal interest points (STIPs) [1] have gained increasing popularity for action recognition since they describe salient points in space and time and have demonstrated strong recognition performance. Nevertheless, spatio-temporal features may fail when the activities become complex since they are unable to capture the global spatial relationships in the subject performing the action [2]. Conversely, graphs are a powerful tool to represent structured objects and as such have been used for action recognition in a recent work from Ta *et al* [3]. Nevertheless, in [3] graphs are directly compared to assess the similarity of two action instances, a procedure that is prone to significant noise. An efficient alternative to the direct comparison of action graphs is offered by graph embedding [4]: in each frame, the graph representing the actor's shape can be converted to a finite set of distances from prototype graphs, and the distance vector then used as

G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 474–482, 2012.

a feature vector with conventional statistical classifiers. Other approaches leveraging on a graphical representation of the actor are based on models akin to Pictorial Structures [5]. Such models were originally proposed for limb motion tracking and require higher resolution imagery to ensure accurate fitting. In all cases, purely structural approaches do not take advantage of the useful information offered by spatio-temporal appearance descriptors.

In this paper, we introduce a novel framework for the fusion of the structural information provided by graph embedding and the spatio-temporal information given by STIP descriptors, thus benefitting from both powerful representations and overcoming their respective limitations. Experiments are performed over the popular dataset KTH [6].

The remainder of this paper is organised as follows. Firstly, in section 2 we define the feature set used in our framework. The proposed approach is then described in section 3. In section 4, we present an experimental evaluation of the proposed method on the KTH action dataset. Finally, conclusions and discussion of future work (section 5).

2 Features

The following section provides a description of the *structural* and *spatio-temporal* features provided by graph embedding and typical descriptors such as those extracted from STIPs, respectively.

2.1 Structural Features

Graphs can represent many patterns very effectively by adjusting the graphs' complexity to that of the patterns. However, their main limitation is that they are computationally cumbersome for pattern analysis. One method of circumventing this problem is that of transforming the graphs into a vector space by means of graph embedding. This section briefly provides an overview of prototype-based graph embedding and then describes its use for incorporating structural information into feature vectors.

Overview of Prototype-Based Graph Embedding
In this work, we avail of the definition of *attributed graph*, noted as $g = (V, E, \alpha, \beta)$ with:

- $V = \{1, 2, ..., M\}$, a set of vertices (nodes),
- $E \subseteq (V \times V)$, a set of edges,
- $\alpha : V \to L_V$, a vertex labeling function, and
- $\beta : E \to L_E$, an edge labeling function.

Vertex and edge labels are restricted to fixed-size tuples, ($L_V = \mathbb{R}^p$, $L_E = \mathbb{R}^q$, p, $q \in \mathbb{N} \cup \{0\}$). When attributed graphs are used to represent objects, the problem of pattern recognition changes to that of graph matching. One of the most widely used methods for error-tolerant graph matching is the graph edit distance (GED), defined as the cost of a transformation "morphing" a given graph into another [7]. GED measures the (dis)similarity of arbitrarily structured and arbitrarily labeled graphs and is

flexible thanks to its ability to cope with any kind of structural errors [7]. The edit transformation is usualy broken up into atomic edit operations which can be of six basic types: insertion, deletion and substitution, for either nodes or edges, and noted as $(e^{i,n}, e^{d,n}, e^{s,n}, e^{i,e}, e^{d,e}, e^{s,e})$. It can be proven that every arbitrary graph can be transformed into another, equally arbitrary graph by applying a finite sequence of edit operations (also called an *edit path*). The distance between the two graphs is defined as the minimum cost amongst all edit paths transforming the first graph into the other. Let $g_i = (V_i, E_i, \alpha_i, \beta_i)$ and $g_j = (V_j, E_j, \alpha_j, \beta_j)$ be a pair of graphs in a set. The graph edit distance of such graphs is formally defined as:

$$d(g_i, g_j) = \min_{(e_1,\ldots,e_k)\in E(g_i,g_j)} \sum_{l=1}^{k} C(e_l) \tag{1}$$

where $E(g_i, g_j)$ denotes the set of edit paths between the two graphs, C denotes the edit cost function and e_l denotes the individual edit operation. Based on (1), the problem of evaluating the structural similarity of two graphs is changed into that of finding a minimum-cost edit path between them. Among the various methods, the *probabilistic graph edit distance* (P-GED) proposed in [8] is capable of automatically inferring the cost function from a training set of manually-paired graphs. P-GED measures the similarity of two graphs by a learned probability, $p(g_i, g_j)$, and defines the dissimilarity measure as: $d(g_i, g_j) = -\log p(g_i, g_j)$. A further advantage of P-GED is its claimed ability to learn from large sets of graphs with huge distortion between samples of the same class, which makes it suitable for application to vision problems [8].

Graph Embedding. In the literature, "graph embedding" refers interchangeably to the embedding of a graph as a whole into a point in vector space, or the embedding of its set of nodes into a set of corresponding points in vector space. In this work, we assume the former meaning, although similar embedding techniques can be applied in the two cases and for other types of non-vectorial objects such as strings or trees [9]. The embedding assumes that a set of objects is given alongside distance values between any two objects in the set. The goal is that of converting the set of objects into a set of points in a vector space of given dimensionality while ensuring certain properties or constraints. Well-known embedding techniques include Laplacian eigenmaps, commute times, symmetric polynomials, amongst others [10], [11], [12]. After the embedding of the initial set of objects, it is also possible to embed new, out-of-sample objects, albeit not always straightforward. An alternative embedding approach is to make use of a given set of "prototype" objects (or prototypes, for short) which can equally embed in-sample and out-of-sample data, in a way not unlike that of eigenvectors in principal component analysis. Let $G = \{g_1, g_2, ..., g_m\}$ be a set of graphs, $P = \{p_1, p_2, ..., p_n\}$ be a set of prototype graphs with $n < m$, and d be a dissimilarity measure. For embedding any graph $g_j \in G$ by way of P, the dissimilarity measure $d_{ji} = d(g_j, p_i)$ of graph g_j to prototype $p_i \in P$ is computed $\forall i$. Then, an n-dimensional vector $(d_{j1}, ..., d_{jn})$ is assembled from all the n dissimilarities. With this procedure, any graph can be individually transformed into a vector of real numbers [13]. Prototype-based embedding is certainly the simplest and fastest embedding approach and for these reasons is adopted hereafter.

Prototype Selection. Selecting informative prototypes from the underlying graph domain plays a vital role in graph embedding [13]. In order to obtain a meaningful as well as class-discriminative vector representation in the embedding space, a set of selected prototypes $P = \{p_1, p_2, ..., p_n\}$ should be adequately distributed over the whole graph domain, at the same time avoiding redundancies in terms of selection of similar graphs [13] ,[14]. Among various prototype selection algorithms [13], [15], [16], the *discriminative prototype selection* method [15] was chosen in this study. This approach select prototypes from a graph set by adequately balancing within-class and between-class scattering.

Structural Features Extraction

The approach used for extracting structural features consists of the following main steps:

1. Use of a modified tracker [17] to extract a bounding box of each actor in each frame, and detection of the scale-invariant feature transform (SIFT) keypoints [18] within such a bounding box by using the software of Vedaldi and Fulkerson [19]. Based on the chosen threshold, this number for the selected dataset (KTH [6]; details provided in section 4) typically varies between 5 and 8. After detection, the location of each SIFT keypoint, (x, y), is expressed relatively to the actor's centroid and employed as a node label for an attributed graph describing the human's shape. In a preliminary study, we found that graphs with only labeled nodes granted comparable accuracy to graphs with both labeled nodes and labeled edges, yet resulted in faster processing. We therefore decided to employ graphs consisting only of labeled nodes (labeled edgeless graphs).
2. Next, in order to identify a prototype set which could lead to meaningful feature vectors in the embedded space, a number of different reference postures was chosen to describe all human shapes in the action dataset. For the dataset at hand (KTH [6]), we arbitrarily chose a set of 16 different reference postures across all human actions (running, walking, boxing, jogging, hand-waving, hand-clapping). Such selected postures should prove adequate for recognising human actions also in any other dataset where the actors are approximately in full view such as UCF Sports [20] and MuHAVi [21]. For training purposes, we manually selected a number of different frames varying in scenario (e.g. outdoor, outdoor with different clothes, indoor), action (e.g. hand waving, hand clapping, jogging) and actor (e.g. person01, person25, person12).
3. Finally, the graph is embedded into the feature vectors by means of P-GED with the prototype set of choice.

2.2 Spatio-Temporal Features

In this paper, in order to establish a fair comparison and focus the scope on the benefits of structural information, we have chosen to adopt the same features - STIP descriptors - of a deservedly much-cited paper from Laptev [1]. STIP descriptors have gained increasing popularity for action recognition since they describe salient points in space and time and do not require a preliminary step of foreground extraction which is generally

regarded as inaccurate. A STIP descriptor consists of the concatenation of a histogram of quantised gradient (HOG) and a histogram of quantised optical flow (HOF) computed over a small spatio-temporal volume of pixels [1]. In this paper, we have used a combination of HOG and HOF for an overall dimensionality of 145. The main difference with [1] is that we do not convert descriptors into codewords; rather, we use each descriptor individually as an observation for our model (details in section 3).

3 Graphical Model

In this paper, we have used the *hidden Markov model with multiple, independent observations (HMM-MIO)* [22], a modified hidden Markov model (HMM) [23] capable of dealing with sequences of observations that include outlier, high-dimensional, and sparse measurements typical of action recognition.

Robustness to outliers is obtained by modelling the observation densities with Student's t distributions [24]. Dimensionality reduction is implemented by using the probabilistic principal component framework [25], and multimodality is taken into account by using a mixture distribution. Finally, modifications to the Baum-Welch algorithm allow for a variable number of observations per frame (single, multiple or none) by the assumption of independence and identical distribution of observations given the state of the HMM. This is a simplifying assumption given that in reality dependencies between these observations may exist. By noting as $O_t \equiv O_t^{1:N_t}$ the set of observations at time t, N_t their number, and Q_t the corresponding hidden state, we define:

$$P(O_t^{1:N_t}|Q_t) \equiv P(O_t^1, ..., O_t^{N_t}|Q_t) = \prod_{n=1}^{N_t} P(O_t^n|Q_t), if N_t > 1 \tag{2}$$

and

$$P(O_t^{1:N_t}|Q_t) \equiv 1, if N_t = 0 \tag{3}$$

Posing $P(O_t^{1:N_t}|Q_t) = 1$ in the case of no observations is equivalent to a missing observation and has neutral effect in the chain evaluation of the HMM-MIO.

In this study, the probability for all the observations in a frame, t, is calculated by the fusion of two likelihoods which model two types of measures:

- Spatio-Temporal Texture or Appearance Observations ($O_{a,t}$) provided by the STIP descriptors: the different numbers of STIP points per frame introduced a scale problem in the resulting probability that is solved in HMM-MIO by means of the following normalization:

$$P_a(O_{at}^{1:N_t}|Q_t) = \sqrt[N_t]{\prod_{n=1}^{N_t} P(O_{a,t}^n|Q_t)} \tag{4}$$

- Structural Observations ($O_{s,t}$) provided by graph embedding: In our experiments, the embedding of a graph with 16 different selected prototypes leads to a 16-dimensional feature vector describing the shape of a single actor in each frame. This feature vector is modelled statistically by likelihood $P(O_{s,t}|Q_t)$.

The combination of the two likelihoods (equation 5) is performed as a weighted sum of weights W_a and W_s, such that $W_a + W_s = 1$.

$$P(O_t|Q_t) = W_a \cdot P_a(O_{a,t}^{1:N_t}|Q_t) + W_s \cdot P(O_{s,t}|Q_t) \tag{5}$$

The graphical model for the modified HMM-MIO can be seen in Figure 1. The generative model is then obtained as the joint probability $P(O_{1:T}, Q_{1:T}|\lambda)$ of a sequence of observations, $O_{1:T} \equiv \{O_1, ..., O_t, ..., O_T\}$, and a sequence of corresponding hidden states, $Q_{1:T} \equiv \{Q_1, ..., Q_t, ..., Q_T\}$.

$$P(O_{1:T}, Q_{1:T}|\lambda) \equiv p(O_1, Q_1, ..., O_t, Q_t, ..., O_T, Q_T|\lambda) \tag{6}$$

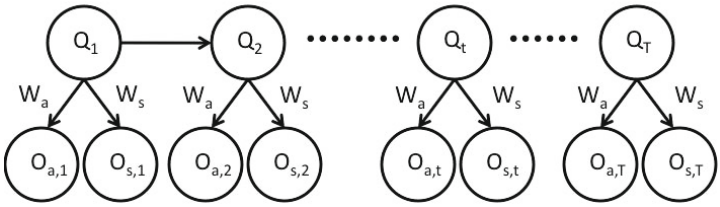

Fig. 1. Modified HMM-MIO (hidden Markov model with multiple, independent observations); O_t are the observations at time t (appearance observations provided by the STIP descriptors, O_a, and the structural observation provided by graph embedding, O_s); Q_t is the corresponding hidden state; W_a and W_s are the two weights for computing the total observation probability $P(O_t|Q_t) = W_a \cdot P_a(O_{a,t}^{1:N_t}|Q_t) + W_s \cdot P(O_{s,t}|Q_t)$; $W_a + W_s = 1$

4 Experiments

This section provides the experimental evaluation of the proposed approach and shows the advantages of combining the structural information provided by graph embedding with the spatio-temporal information provided by STIPs. As dataset, we have chosen the KTH human action dataset containing 2,391 video sequences (from 25 different actors) and acquisition conditions, inclusive of four different scenarios and mild camera movements. The action classes include walking, jogging, running, boxing, hand waving and hand clapping [6]. Although KTH is becoming saturated in recent years with results reporting high accuracies, it still offers the widest platform for comparison with previous work [26]. For accuracy evaluation, we have used the evaluation procedure proposed by Schuldt *et al.* in [6]. In this procedure, all sequences are divided into three sets with respect to the actors: training (8 actors), validation (8 actors) and test (9 actors). Each classifier is then tuned using the first two sets (training and validation sets), and the accuracy on the test set is measured "blindly" by using the parameters selected on the validation set, without any further tuning. In order to assess the individual contribution of the features and show the advantages of the proposed fusion, we have conducted experiments with different weights (Table 1). A value of $(W_a, W_s) = (1, 0)$ means that only appearance features are used, whereas structural features are solely utilised

Table 1. Accuracy (%) of our approach over the KTH dataset with variable weights over the appearance and structural components.

W_a	W_s	Test accuracy (%)
1.0	0.0	*85.7* [22]
0.9	0.1	85.9
0.8	0.2	86.8
0.7	0.3	87.9
0.6	0.4	88.9
0.5	**0.5**	**89.8**
0.4	0.6	87.9
0.3	0.7	85.8
0.2	0.8	82.2
0.1	0.9	77.9
0.0	1.0	*48.7* [4]

when $(W_a, W_s) = (0, 1)$. As shown by Table 1, recognition accuracy is significantly improved by combining the structural information with the spatio-temporal features, reaching its maximum when $(W_a, W_s) = (0.5, 0.5)$.

To position our work properly, it is very important to state that current results on KTH are well in excess of 90% accuracy [27]. The goal of our paper is not that of proposing a more accurate action recognition method; rather, assessing the fusion of structural information with spatio-temporal features in a significant classification exercise. As for what action recognition is concerned, we have gathered empirical evidence that the graphs built by using SIFT keypoints as their nodes are rather unstable and noisy, and we are working on the use of graph-cut techniques to substantially improve nodes' extraction [28]. However, we believe that the work conducted to date already provides evidence that the fusion of structural information obtained by graph embedding with spatio-temporal information provided by STIPS is capable of encoding the human action to a significant extent.

5 Conclusions and Future Work

In this paper, we have presented a novel approach for human action recognition based on the fusion of structural and spatio-temporal information. To this aim, the structural information provided by graph embedding and the local spatio-temporal information provided by STIP descriptors are jointly modelled by a modified hidden Markov model with multiple, independent observations (HMM-MIO) [22]. Although our approach does not yet outperform the state-of-the-art accuracy, it shows that structural and spatio-temporal features can be fused constructively to obtain higher accuracy than from either separately. In the near future, we plan to further investigate other keypoint sets to improve the stability of the graph-based representation along the frame sequence and extend our study to other challenging action datasets.

Acknowledgments. The authors wish to thank the Australian Research Council and its industry partners that have partially supported this work under the Linkage Project funding scheme - grant LP 0990135 “Airport of Future“.

References

1. Laptev, I.: On space-time interest points. International Journal of Computer Vision 64(2), 107–123 (2005)
2. Niebles, J., Chen, C.W., Fei-Fei, L.: Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 392–405. Springer, Heidelberg (2010)
3. Ta, A.-P., Wolf, C., Lavoue, G., Baskurt, A.: Recognizing and localizing individual activities through graph matching, pp. 196–203. IEEE Computer Society, Los Alamitos (2010)
4. Borzeshi, E.Z., Xu, R.Y.D., Piccardi, M.: Automatic Human Action Recognition in Videos by Graph Embedding. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part II. LNCS, vol. 6979, pp. 19–28. Springer, Heidelberg (2011)
5. Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. IEEE Transactions on Computers 22(1), 67–92 (1973)
6. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 3 (2004)
7. Gao, X., Xiao, B., Tao, D., Li, X.: A survey of graph edit distance. Pattern Analysis & Applications 13(1), 113–129 (2010)
8. Neuhaus, M., Bunke, H.: Automatic learning of cost functions for graph edit distance. Information Sciences 177(1), 239–247 (2007)
9. Rieck, K., Laskov, P.: Linear-Time Computation of Similarity Measures for Sequential Data. Journal of Machine Learning Research 9, 23–48 (2007)
10. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation 15(6), 1373–1396 (2003)
11. Qiu, H., Hancock, E.R.: Clustering and embedding using commute times. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(11), 1873–1890 (2007)
12. Wilson, R.C., Hancock, E.R., Luo, B.: Pattern vectors from algebraic graph theory. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1112–1124 (2005)
13. Riesen, K., Neuhaus, M., Bunke, H.: Graph Embedding in Vector Spaces by Means of Prototype Selection. In: Escolano, F., Vento, M. (eds.) GbRPR. LNCS, vol. 4538, pp. 383–393. Springer, Heidelberg (2007)
14. Hjaltason, G.R., Samet, H.: Properties of embedding methods for similarity searching in metric spaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(5), 530–549 (2003)
15. Borzeshi, E.Z., Piccardi, M., Xu, R.Y.D.: A discriminative prototype selection approach for graph embedding in human action recognition. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1295–1301. IEEE (2011)
16. Riesen, K., Bunke, H.: Graph classification by means of Lipschitz embedding. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39(6), 1472–1483 (2009)
17. Chen, T.P., Haussecker, H., Bovyrin, A., Belenov, R., Rodyushkin, K., Kuranov, A., Eruhimov, V.: Computer vision workload analysis: case study of video surveillance systems. Intel Technology Journal 9(2), 109–118 (2005)
18. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)

19. Vedaldi, A., Fulkerson, B.: Vlfeat: An open and portable library of computer vision algorithms. In: Proceedings of the International Conference on Multimedia, pp. 1469–1472. ACM (2010)
20. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
21. Singh, S., Velastin, S.A., Ragheb, H.: Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 48–55. IEEE (2010)
22. Concha, O.P., Xu, D., Yi, R., Moghaddam, Z., Piccardi, M.: Hmm-mio: an enhanced hidden markov model for action recognition. In: 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 62–69. IEEE (2011)
23. Rabiner, L., Juang, B.: An introduction to hidden markov models. IEEE ASSP Magazine 3(1), 4–16 (1986)
24. Liu, C., Rubin, D.B.: Ml estimation of the t distribution using em and its extensions, ecm and ecme. Statistica Sinica 5(1), 19–39 (1995)
25. Archambeau, C., Delannay, N., Verleysen, M.: Mixtures of robust probabilistic principal component analyzers. Neurocomputing 71(7), 1274–1282 (2008)
26. Gao, Z., Chen, M., Hauptmann, A., Cai, A.: Comparing Evaluation Protocols on the KTH Dataset. In: Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A. (eds.) HBU 2010. LNCS, vol. 6219, pp. 88–100. Springer, Heidelberg (2010)
27. Guo, K., Ishwar, P., Konrad, J.: Action recognition using sparse representation on covariance manifolds of optical flow. In: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 188–195. IEEE (2010)
28. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics (TOG) 23, 309–314 (2004)