Human interaction recognition based on the co-occurence of visual words

K. Nour el houda SLIMANI† Yannick BENEZETH†† Feriel SOUAMI†

† USTHB, LRIA Laboratory
BP 32 El Alia Bab Ezzouar,
Algiers 16111, Algeria
[nslimani;fsouami]@usthb.dz

†† Université de Bourgogne
LE2I, UMR CNRS 6306
21000 Dijon cedex France
yannick.benezeth@u-bourgogne.fr

Abstract

This paper describes a novel methodology for automated recognition of high-level activities. A key aspect of our framework relies on the concept of co-occurring visual words for describing interactions between several persons. Motivated by the numerous success of human activity recognition methods using bag-of-words, this paradigm is extended. A 3-D XYT spatio-temporal volume is generated for each interacting person and a set of visual words is extracted to represent his activity. The interaction is then represented by the frequency of co-occurring visual words between persons. For our experiments, we used the UT-interaction dataset which contains several complex human-human interactions.

1. Introduction

Over the past several years, human activity recognition has attracted the attention of the computer vision community. Understanding actions in videos is a problem that has been studied extensively and successful results have been obtained for the recognition of quite simple actions [20]. On the other hand, the recognition of more complex actions that may involve several persons is still an active research topic.

The method proposed in this paper is inspired by the bag-of-words paradigm which is one of the most popular method in the field of textual information retrieval and object recognition in images. Various methodologies based on this paradigm have been developed for activity recognition in videos, *e.g.* in [12, 16, 6, 13]. The action recognition is achieved using an unordered set of spatio-temporal features, called visual words. Such features capture local motion events in video. These visual words are usually combined with a state of the art machine learning techniques such as a Boosting or Support Vector Machines classifiers to recognize human actions.

Encoding the spatio-temporal structure of visual words is of great importance for the recognition of human interaction over video sequences. Complex actions, such as interactions between individuals, are composed of a set of temporaly ordered elementary actions performed by the different persons involved in the interaction. The temporal sequence is not considered in the original bag-of-words paradigm. To cope with this limitation, we present a spatio-temporal representation for interaction recognition based on the co-occurrence of visual words. The co-occurrence matrix is a non parametric model for pair-wise feature distribution and is used to effictively model the frequency of co-occurrence of visual words.

An overview of the proposed methodology is presented in figure 1. A 3-D spatio-temporal volume is generated for each interacting person and a set of visual words is extracted to represent his activity. The interaction between persons is then represented by the frequency of co-occurring visual words, *i.e.* the number of times visual words occur simoustaneously for each person involved in the interaction.

The paper is organized as follow. First, related works are presented in section 2, then the human interaction representation is described in section 3 followed by experimental results in section 4.

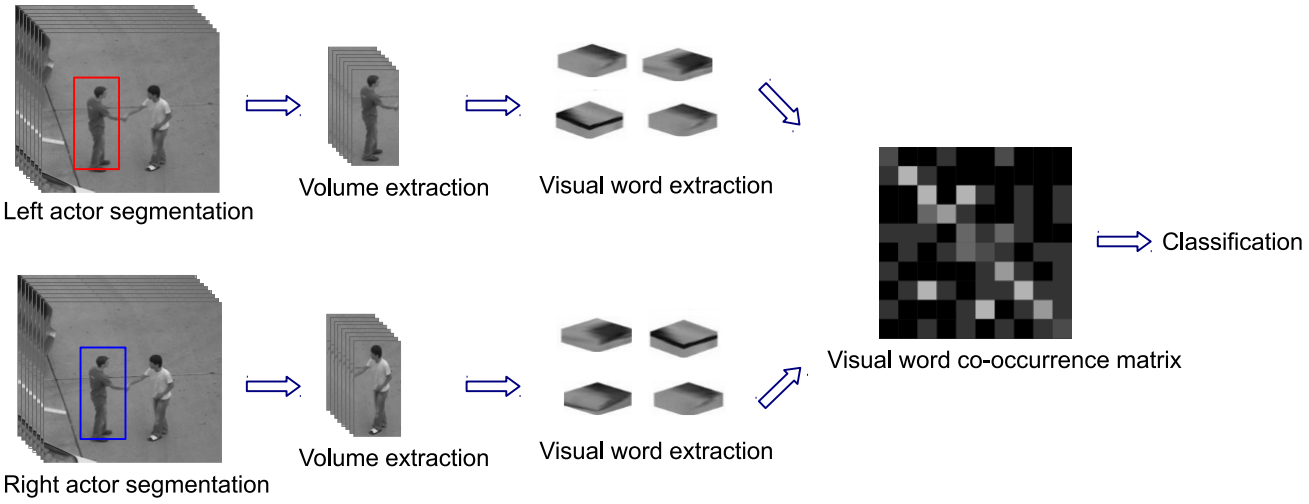

Figure 1. Overview of the proposed approach.

2. Related work

Much of the early work in action recognition was tested on relatively simple single person uniform background sequence. Most datasets for human action recognition, such as the *KTH* [20] or the *Weizmann* [3] datasets, provide samples for only a few action classes recorded in controlled and simplified settings. Recently, more complex datasets have been proposed with realistic video samples, such as the *Hollywood* dataset [10] or more complex actions and interactions such as the *UT-interaction* dataset [18].

There are several existing surveys within the area of high-level human activity recognition [22, 15, 4, 1]. We present in this section the most common techniques.

Local features are a popular way for representing actions in videos. They achieve state-of-the-art results for action classification when combined with a bag-of-words representation [21]. However, by exploiting correlations in space and time between these local features, actions can be modeled more effectively. Unlike the approaches following the bag-of-words paradigm, some approaches attempt to model spatio-temporal distribution of the local features for better recognition of actions [1, 5, 24, 9, 7, 14, 23]. Savarese *et al.* [19] proposed a methodology to capture spatio-temporal proximity information among local features. For each action video, they measured co-occurrence patterns in a local 3-D region. Liu and Shah [11] also considered correlations among local features.

Laptev *et al.* [10] constructed spatio-temporal histograms by dividing an entire space-time volume into several grids. The method roughly measures how local descriptors are distributed in the 3-D XYT space by analyzing which feature falls into which cell of the grid.

Ryoo and Aggarwal [17] introduced the spatio-temporal relationship match, called *STR match*, which explicitly considers spatial and temporal relationships among local features. Their method measures structural similarity between two videos by computing pairwise spatio-temporal relations among local features. Their system not only classified simple actions (*i.e.* those from the *KTH* dataset), but also recognized interaction-level activities (*e.g.* hand-shaking and pushing) from continuous videos.

In this paper, we are also interested in exploiting correlations in space and time between local features. We propose a non-parametric representation for pairwise feature distribution to explicitely consider the simultaneous occurrence of visual words between the persons involved in an interaction. The interaction representation is based on the frequency of co-occurring visual words between persons.

3. Human interaction representation

In this section, we introduce the proposed representation of human interactions. An overview of our

approach is described in figure 1. First, a 3-D XYT spatio-temporal volume is extracted for each interacting person and the framework of bag-of-words is used to characterize the activity. Then, as we are interested in the recognition of complex activities involving at least two persons, we use the frequency of co-occurring visual words to get a compact representation of the activity. Interactions that we recognize in this paper do not involve more than two persons, but it may be noted that our method can be easily generalized to interactions involving more than two persons.

3.1. Low-level action representation

The first step of the proposed approach is to detect and segment the persons in the video in order to create a 3-D XYT spatio-temporal volume for each of these persons. The activity of each person is then characterize with a set of local features extracted from this 3D volume. It may be noted that it is possible to use any human detection and tracking methods [2] as well as any local features suitable for spatio-temporal volumes such as 3D-SIFT [21], 3D-HOG [8] or 3D-LBP [25] to name a few. In order to decrease the sensitivity of our method to tracking errors, we have decided to manually extract the location of each person in the video. We have also chose to use the 3D-SIFT features [21]. This descriptor has been applied successfully in various tasks such as action classification.

Once local features are extracted for all videos, we cluster them into multiple visual words using *k-means* algorithm in order to define a visual codebook. Visual-words are then defined by the centroids of the clusters.

Thus, each spatio-temporal volume V is defined by a set of visual-word tuples:

$$V = \{\langle x_0, y_0, t_0, w_0\rangle, \ldots, \langle x_n, y_n, t_n, w_n\rangle\}, \quad (1)$$

where each tuple $\langle x_i, y_i, t_i, w_i\rangle$ represents respectively the spatial position, the time and the word index of the i^{th} interest point. n denotes the volume's total number of interest points. $w_i \in \{1, \ldots, k\}$ with k the number of visual words.

In the original bag-of-words paradigm, the video sequence is represented as a histogram of visual-words frequencies. The easiest way to handle cases where there are several persons in the video is to merge the visual words from the different persons and to build a single histogram of visual-word frequencies. This method is called *BoW* in the experimental results section 4. In the following, we introduce the concept of visual-words co-occurrence to describe the multi-person actions.

3.2. Multi-person action representation

To simplify notations in the following paragraphs, we will consider that there are only two interacting persons in the video. Consequently, the video is composed of only two spatio temporal volumes $V^{(1)}$ and $V^{(2)}$.

Let $\vartheta_i^{(p)}$ be the i^{th} visual-word tuple of the p^{th} volume $\langle x_i^{(p)}, y_i^{(p)}, t_i^{(p)}, w_i^{(p)}\rangle$ with $p \in \{1, 2\}$. We consider that $\vartheta_i^{(1)}$ and $\vartheta_j^{(2)}$ co-occur whenever $|t_i^{(1)} - t_j^{(2)}| < \tau$. The visual-word co-occurrence matrix $\mathcal{C}$ is defined with

$$\mathcal{C}_{i,j} = \frac{1}{M} \sum_{t=0}^{T-1} \delta(\vartheta_i^{(1)}, \vartheta_j^{(2)}) \quad (2)$$

where $\delta(\vartheta_i^{(1)}, \vartheta_j^{(2)}) = 1$ if $\vartheta_i^{(1)}$ and $\vartheta_j^{(2)}$ co-occur and 0 otherwise, M is a normalization term and T the number of frames in the volumes. A visual-word co-occurrence matrix example is presented in figure 2.

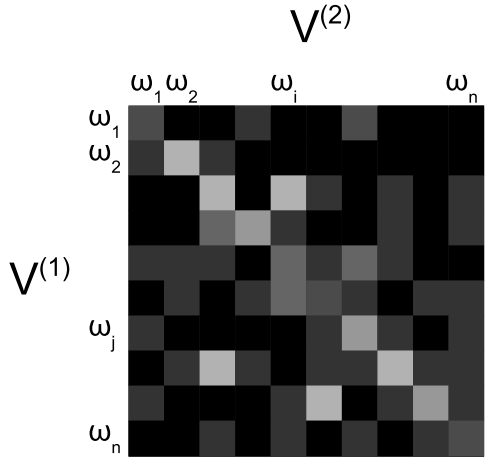

Figure 2. Example of visual-word co-occurrence matrix extracted from two spatio-temporal volumes (V1 and V2).

Note that we later modify the co-occurrence matrix in order to be robust to the relative position of the persons. We use $\mathcal{C}'_{i,j} = \mathcal{C}_{i,j} + \mathcal{C}_{j,i}$ and we finally keep the

upper triangular part of the matrix $\mathcal{C}'$ matrix for representing the interaction. These features can later be used for classification using regular machine learning classifiers.

Interactions that we recognize in this paper come from the *UT-interaction* dataset and do not involve more than two persons. But it may be noted that our method can be easily generalized to interactions involving more than two persons. If the video is composed of m volumes, with $m > 2$, it is possible to consider a m-dimensional co-occurrence matrix $\mathcal{C}$ where each element of the matrix represent the co-occurrence of m visual words from the m volumes.

4. Experimental results

This section is committed to assess the performance of our framework and to present the preliminary results regarding the recognition of high level human interactions. The experiments were carried out using the *UT-interaction* dataset [18] which contains videos recorded under a realistic surveillance environment. The *UT-interaction* dataset contains six classes of human-human interactions: *shake-hands*, *point*, *hug*, *push*, *kick* and *punch*. Illustrations of these interactions are presented in figure 3.

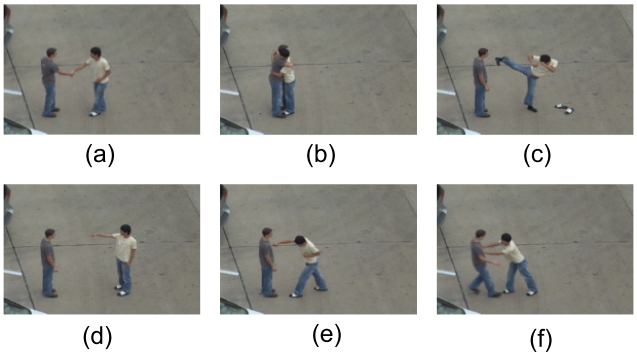

Figure 3. Examples of interactions in the UT-Interaction dataset. (a) hand-shaking; (b) hugging; (c) kicking; (d) pointing; (e) punching; (f) pushing.

The dataset is composed of 20 video sequences divided into two sets. The set #1 is composed of 10 video sequences taken on a parking lot with slightly different zoom rate, and their backgrounds are mostly static with little camera jitter. Whereas the set #2 (*i.e.* the other 10 sequences) are taken on a lawn in a windy day. Background is moving slightly (*e.g.* with trees shaken by the wind), and these videos contain more camera jitters.These 120 video segments representing six occurrences of interaction's class are proposed from the 20 sequences and are used for the classification evaluation.

Table 1. Per clip activity classication performance measured as ROC Area (in %) on the *UT-Interaction Dataset* obtained for the Set 1.

	Shake	Hug	Kick	Punch	Push
ROC A.	0.724	0.726	0.426	0.564	0.577

Table 2. Per clip activity classication performance measured as ROC Area (in %) on the *UT-Interaction Dataset* obtained for the Set 2.

	Shake	Hug	Kick	Punch	Push
ROC A.	1.000	1.000	0.591	0.739	0.500

Table 3. Comparison of our method with the regular *BoW* approach. Average accuracies on the *UT-Interaction Dataset*.

Methods	Set 1	Set 2
BoW	58.20	48.30
Our method	40.63	66.67

Similarly to [12], we temporally extract key frames from the original video's middle one third frames with five frames as the sampling interval. As explained above, in order to decrease the sensitivity of our method to tracking errors, we have decided to manually extract the location of each person in the video. We have also chose to train our human interaction classifiers using the 3D-SIFT features [21] as described above. The experiments conducted shows that the selection of the number of visual words is critical. Using a bigger number of visual words does not bring better performance because of the size of the feature vector. Therefore, the use of dimension reduction techniques, such as PCA, should be of a great importance because it would allow the use of a more descriptive vocabulary. In the following, we use 150 visual words in the vocabulary and the threshold τ is set to 1 for defining the co-occurrences. It can be noted that any state of the art classifier can be used with our interaction representation. In our experiments, for Set 1, we use a simple locally weighted learning with k-nearest neighbor classifier and Euclidean Distance as distance func-

	Shake	Hug	Kick	Punch	Push
shake	0.50	0.00	0.00	0.33	0.16
Hug	0.14	0.75	0.14	0.00	0.00
kick	0.0	0.28	0.14	0.28	0.28
Punch	0.16	0.00	0.33	0.33	0.16
Push	0.16	0.16	0.16	0.16	0.33

	Shake	Hug	Kick	Punch	Push
shake	1.00	0.00	0.00	0.00	0.00
Hug	0.00	1.00	0.00	0.00	0.00
kick	0.00	0.00	0.50	0.50	0.00
Punch	0.00	0.00	0.25	0.75	0.00
Push	0.00	0.00	0.50	0.50	0.00

Table 4. Confusion matrices of per-clip classification result on UT-Interaction dataset. Horizontal rows are ground truth and vertical columns are predictions.

tion for implementing the nearest neighbor search. For the Set 2, we use a SVM classifier with a polynomial kernel. Furthermore, we use a 5-fold cross-validation experimental setting to evaluate our method. It is noteworthy that in our experiment, we have used 32 instances (32 from 60 videos) and we have ommited the class *Point* because it can be considered as an action and not an interaction.

We present in tables 1 and 2 the results obtained for the sets 1 and 2 of the UT-Interaction dataset. We present the per clip activity classication performance as ROC Area (in %). More detailled results are presented in table 4. Confusion matrices of both the two sets in the UT-Interaction datsets are shown for our method. It is possible to observe some confusion between the activities *push*, *punch* and *kick* on set 2 as those interactions are slightly similar in both appearance and motion.

Then, we present in table 3 a comparison of our method with the regular Bag-of-words (*BoW*) approach. For this method we have extracted 3D SIFT features for each person's volume involved in a given interaction and then we have merged the two histogram into a single histogram of visual words to define an interaction. The results in table 3 are presented in terms of average accuracies. It is possible to observe that our method performs better on the Set 2 but not on the Set 1. These encouraging results suggest that a more appropriate management of the size of our descriptors should allow to use a more descriptive vocabulary and thus further improve the results.

5. Conclusion

We have presented in this paper a representation of actions with several persons based on the concept of co-occurrence of visual words for describing interactions. The paradigm of bag-of-words for action recognition is extended and the multi-person actions are then represented by the frequency of co-occurring visual words, *i.e.* the number of times visual words occur simoustaneously between the persons involved in the interaction. We have presented experimental results on the *UT-interaction* dataset using the 3-D SIFT local features and a SVM classification. In the future, we plan to evaluate more deeply the robustness of the proposed method to the various parameters, to experiment the extension to videos with more than two persons.

References

[1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):1–43, 2011. 2

[2] Y. Benezeth, B. Emile, H. Laurent, and C. Rosenberger. Vision-based system for human detection and tracking in indoor environment. *International Journal of Social Robotics*, 2(1):41–52, 2010. 3

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *IEEE International Conference on Computer Vision*, volume 2, pages 1395–1402, 2005. 2

[4] J. Candamo, M. Shreve, D. Goldgof, D. Sapper, and R. Kasturi. Understanding transit scenes: A survey on human behavior-recognition algorithms. *IEEE Trans. on Intelligent Transportation Systems*, 11(1):206–224, 2010. 2

[5] O. Chomat and J. Crowley. Probabilistic recognition of activity using Local appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1999. 2

[6] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *British Machine Vision Conference*, volume 2, page 7, 2010. 1

[7] P. Dollar, P. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE int. Workshop on Visual Surveillance*

and Performance Evaluation of Tracking and Surveillance, pages 65–72, 2005. 2

[8] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, pages 995–1004, 2008. 3

[9] I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE International Conference on Computer Vision*, pages 432–439, 2003. 2

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE int. conf. on Computer Vision and Pattern Recognition*, 2008. 2

[11] J. Liu and M. Shah. Learning human action via information maximization. In *IEEE int. conf. on Computer Vision and Pattern Recognition*, 2008. 2

[12] L. Meng, L. Qing, P. Yang, J. Miao, X. Chen, and D. Metaxas. Activity recognition based on semantic spatial relation. In *IEEE International Conference on Pattern Recognition*, 2012. 1, 4

[13] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3), 2008. 1

[14] J. Niebles, H. Wang, and F.-F. L. Unsupervised learning of human action categories using spatial-temporal words. In *British Machine Vision Conference*, pages 1249–1258, 2006. 2

[15] P. Oluwatoyin and K. Wang. Video-based abnormal human behavior recognition - a review. *IEEE Trans. on Systems, Man, and Cybernetics, Part C*, 42(6):865–878, 2012. 2

[16] M. H. P. Matikainen and R. Sukthankarl. Representing pairwise spatial and temporal relations for action recognition. European Conference on Computer Vision, 2010. 1

[17] M. Ryoo and J. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *IEEE International Conference on Computer Vision*, pages 1593–1600, 2009. 2

[18] M. S. Ryoo and J. K. Aggarwal. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). *IEEE International Conference on Pattern Recognition Workshops*, 2010. 2, 4

[19] S. Savarese, A. DelPozo, J. Niebles, and L. Fei-Fei. Spatial-temporal correlatons for unsupervised action classification. In *IEEE Workshop on Motion and Video Computing*, pages 1–8, 2008. 2

[20] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *IEEE International Conference on Pattern Recognition*, pages 32–36. 1, 2

[21] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM Multimedia*, 2007. 2, 3, 4

[22] P. K. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Trans. on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008. 2

[23] A. Yilmaz and M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *IEEE International Conference on Computer Vision*, pages 150–157, 2005. 2

[24] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *IEEE int. conf. on Computer Vision and Pattern Recognition*, 2001. 2

[25] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. 3