Novel Multi-Feature Bag-of-Words Descriptor via Subspace Random Projection for Efficient Human-Action Recognition

Ana I. Maqueda, Arturo Ruano, Carlos R. del-Blanco, Pablo Carballeira,
Fernando Jaureguizar, and Narciso García

Abstract

Human-action recognition through local spatio-temporal features have been widely applied because of their simplicity and its reasonable computational complexity. The most common method to represent such features is the well-known Bag-of-Words approach, which turns a Multiple-Instance Learning problem into a supervised learning one, which can be addressed by a standard classifier. In this paper, a learning framework for human-action recognition that follows the previous strategy is presented. First, spatio-temporal features are detected. Second, they are described by HOG-HOF descriptors, and then represented by a Bag of Words approach to create a feature vector representation. The resulting high dimensional features are reduced by means of a subspace-random-projection technique that is able to retain almost all the original information. Lastly, the reduced feature vectors are delivered to a classifier called Citation K-Nearest Neighborhood, especially adapted to Multiple-Instance Learning frameworks. Excellent results have been obtained, outperforming other state-of-the art approaches in a public database.

1. Introduction

In the last years, human-action recognition has undergone an increasing popularity for its applications in many visual-based systems, such as video surveillance [7], Human-Computer Interaction [21], sports video analysis, and video retrieval [5]. Moreover, a new wave of applications is expected to come due to its huge impact in the society. For example, activity monitoring of elderly and disabled people, augmented reality applications that fuse the world, the mankind knowledge, and virtual elements to give support to education and training activities to improve user experience, and in general to reach a new level of interaction with the surrounding world. However, the recognition of human actions from color imagery is a challenging task due to the articulate nature of the human body and its complex dynamics, cluttered backgrounds, people occlusions, varying illumination conditions, and the huge volume of data related with the search and recognition of spatio-temporal patterns. In addition, the same action performed by different people (and even by the same person) can be dramatically different in both appearance and dynamics. Other source of problems is that human actions can be very ambiguous due to the interaction with other people and other objects.

In order to deal with these problems, a wide range of methodologies have been proposed. Some of them are based on modeling human actions defining interaction models between humans and the surrounding objects [19], [18]. These techniques rely on human detection, object detection, and tracking algorithms to compute human and object trajectories that are jointly used to represent the actions to be recognized. Despite their promising results, the training and settings of such systems is highly complex and prone to errors when a subsystem produce failures (for example, detection or tracking errors).

Other methodologies include body shape, in form of silhouettes or poses, as visual cues [4] to perform the action recognition. This representation is simpler and easier to compute, however silhouettes and poses can be difficult to correctly acquire because of occlusions, cluttered backgrounds or complex movements.

In the last years, local spatio-temporal features have become very popular for action recognition since they do not depend on detectors, trackers, and/or silhouette and pose extractors. The general procedure is as follows. First, spatio-temporal interest points (STIP) are detected, and then they are described by either appearance and motion descriptors [15], or by means of motion trajectories [2]. These methods tend to be more robust against noise, scale changes, cluttered backgrounds and occlusions. And, they are typically used together with matching algorithms or machine learning techniques, such as K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) classifiers.

In this paper, a new learning framework for human-action recognition is proposed, which is based on a novel description strategy based on subspace Random Projection (RP), and a Multiple-Instance Learning (MIL) approach to carry out the recognition task. It can be seen in Figure 1. In particular, the proposed framework extracts the structure scene by computing STIP through the 3D Harris detector [13]. Then, the structure scene is described by combining three stages. The first one consists on a multiple-feature description using Histograms of Oriented Gradients (HOG), and Histograms of Optical Flow (HOF) [14]. The second stage applies a Bag-of-Words (BoW) strategy to generate a compact feature vector representation from the bag/cloud of previously computed HOG-HOF descriptors. For this purpose, a visual dictionary is created by clustering the training HOG-HOF descriptors, which is used to compute a histogram of visual words, i.e. the feature vector that describes the human action. The last stage reduces the dimensionality of the resulting histograms using a technique based on subspace random projection, which has the great advantage of being independent of the underlying data structure in the feature space, unlike other methods such Principal Component Analysis (PCA) [11]. Finally, the reduced feature vectors are classified by means of the Citation K-Nearest Neighbor (C-KNN) algorithm [27] to recognize the performed human action, following a *one-against-one* strategy. The proposed human recognition algorithm has been validated on the publicly available TV Human Interaction database (TVHI) [17], and compared with other state-of-the-art algorithms.

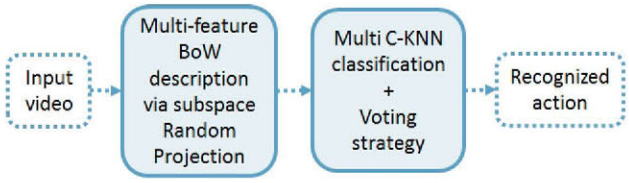

Figure 1. Proposed learning framework for human-action recognition.

The rest of the paper is structured as follows. Section 2 summarizes the recent work in human activity recognition, in particular, those based on local spatio-temporal features. Section 3 describes the proposed feature description. Section 4 explains the employed technique for classification. Section 5 presents the experimental results. And finally, the conclusions are drawn in Section 7.

2. Related work

Local spatio-temporal features usually maximize saliency functions to select spatio-temporal locations and scales that can be useful in the characterization of human action. In order to extract such features, STIP are first detected, and then described. The two main mechanisms to describe STIP are based on motion trajectories and feature descriptors [6]. Regarding the STIP descriptors based on motion trajectories, the KLT feature tracker, proposed by Kanade *et al.* [23], is one of the most employed in a sparse distribution of STIP. Uemura *et al.* [24] extracted a large number of interest points from every frame by using multiple detectors, and described them by using the SIFT descriptor. Finally, they applied the KLT tracker to obtain a trajectory-based description. Messing *et al.* [16] tracked 3D Harris interest points with the KLT tracker. Alternatively, Sun *et al.* [22] modeled spatio-temporal contextual information by matching SIFT descriptors between two consecutive frames, creating a map of sparse motion vectors.

Motion trajectories can be also obtained from densely sampled STIP, as Wang *et al.* introduced in [25]. Interest points were sampled at uniform intervals in space and time and tracked using optical flow fields, obtaining the so-called dense trajectories. Based also on dense trajectories, Jiang *et al.* [9] used global and local interest points to distinguish human-action motion from camera movements. Once the trajectories are obtained, they are characterized using combinations of HOG, HOF, and Motion Boundary Histogram (MBH) descriptors.

On the other hand, the STIP descriptors based on feature descriptors continue showing promising results for action recognition, proving to be robust against scale changes, spatio-temporal shifts, cluttered backgrounds, and multiple motions in the scene. Feature descriptors characterize an STIP neighborhood by extracting both appearance and motion features. Most of them have been just extended from existing image descriptors, such as 3D SIFT [20], and 3D HOG [12]. For example, Chuanzhen *et al.* [15] used the 3D SURF descriptor to represent the local region of interest points, and Zhang *et al.* [29] combined several descriptors such as HOG, HOF, and MBH with motion trajectories.

After local STIP extraction and description, a set of disordered features are obtained per video sequence. In order to create a compact feature vector representation from them, the most common technique is the well-known Bag-of-Words (BoW) approach. This way, a histogram of visual words per video is computed (i.e. the desired feature vector representation), which can be now used by supervised machine learning methods (SVMs, neural networks, random forest, etc.) to perform the human action recognition.

3. Multi-feature Bag-of-Words descriptor based on Subspace Random Projection

The proposed feature descriptor for human-action recognition is shown in Fig. 2, and it is composed by two stages: 1) structure-scene extraction, and 2) structure-scene

description.

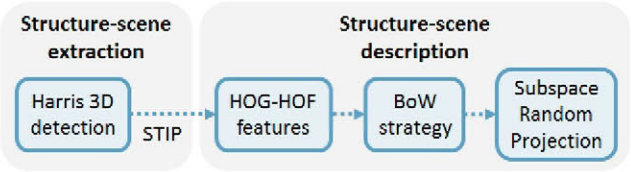

Figure 2. Proposed feature descriptor for human-action recognition.

The structure-scene extraction consists on localizing those spatio-temporal interest points (STIP) that characterize the human actions in the scene. To that end, a variation of the 3D Harris detector [13] is used, which combines multiple spatial and temporal scales, σ_i^2 and τ_j^2 respectively, for the detection process [14], instead of performing scale selection. An example of the STIP detection step is shown in Fig. 3.

The second step in turn consists on three steps: a HOG-HOF description, a Bag-of-Words representation, and a dimensionality reduction based on Subspace Random Projection (RP). They are explained in detail below.

3.1. HOG-HOF description

Once STIP are detected, it is necessary to characterize them. In particular, HOG and HOF histograms are computed in the spatio-temporal neighborhood of each STIP to describe their appearance by means of gradient information, and their motion by means of optical flow, respectively. The size of each spatio-temporal neighborhood $(\Delta_x, \Delta_y, \Delta_t)$ is related to the scales used in the STIP detection as follows: $\Delta_x = \Delta_y = 2k\sigma$ and $\Delta_t = 2k\tau$. Each spatio-temporal neighborhood is further subdivided into a grid with $n_x \times n_y \times n_t = 3 \times 3 \times 2$ spatio-temporal blocks. For each block, a 4-bin HOG based descriptor and 5-bin HOF based descriptor are computed. The resulting HOG and HOF descriptors from all the neighborhood blocks are concatenated into two features vectors, one with 72 elements and the other with 90 elements, respectively. Finally, both feature vectors are again concatenated in a 162-element vector that fully characterizes the corresponding STIP.

3.2. Bag-of-Words representation

To compact the set of disordered HOG-HOF descriptors computed in the previous stage in a feature vector representation, a BoW strategy is used. This is commonly used when one class/entity can be represented by a *bag* of *instances*. In this case, the video is represented by a bag of HOG-HOF descriptors. However, not all instances are necessarily relevant, since there might be instances inside a bag that do not contain any relevant information about one specific class, or worse, some instances are more representative of other classes, providing confusing information. This is the case of HOG-HOF descriptors coming from background regions or common to several human actions. These circumstances are typical from a semi-supervised learning where a class label can be assigned to every bag, but it is not possible to assign individual labels to the instances inside the bag. This classification problem is known as Multiple-Instance Learning (MIL).

The BoW strategy creates a visual dictionary that maps each bag into a compact feature vector that summarizes the information of the whole bag. This mapping transforms the original MIL problem into a standard supervised learning problem, where each BoW-based feature vector has an associated label that any standard classifier, such as AdaBoost, Neural Networks, and SVM, can be used for training and test. In order to compute the visual dictionary, the K-means algorithm is used to cluster the space of HOG-HOF descriptors computed from the training set. The resulting cluster centroids represent the visual words of the dictionary. Then, a new bag of HOG-HOF descriptors from a testing video sequence is mapped to a histogram of visual words (a feature vector) by finding the nearest visual word (cluster centroid) to every HOG-HOF descriptor according to the Euclidean distance. Thus, each bin of the histogram, which represents a cluster centroid, contains the number of HOG-HOF descriptors that belong to that cluster. Finally, an L2-normalization is applied to the histogram of visual words. As a result, a compact and normalized feature vector is obtained representing the video sequence information.

3.3. Random-projection-based reduction

To obtain a high discrimination capability, large visual dictionaries are used, which generates in turn high dimensional feature vectors. Therefore, the length of the feature vectors computed in the previous section is proposed to be reduced. Dimensionality reduction techniques, such as Principal Component Analysis (PCA) [11] and Random Projections (RP) can improve the classification performance by eliminating redundancy [26]. On the one hand, PCA remains popular because of its simplicty, however it suffers from a number of weak points, such as its implicit assumption of Gaussian distributions, its restriction to orthogonal linear combinations. On the other hand, random-projections theory allows to substantially reduce the dimensionality of a problem while still preserving almost all the data structure of problem. It has shown good results in several applications, such as face recognition [8], and image and text processing [3].

By using RP, each n-dimensional feature vector is treated as a single point in $\mathbb{R}^n$ (n being large). Then, a set of N vectors can be thought of as a point cloud in an $\mathbb{R}^n$ space. This point cloud can be projected into a low-dimensional subspace that preserves its geometrical structure. This is

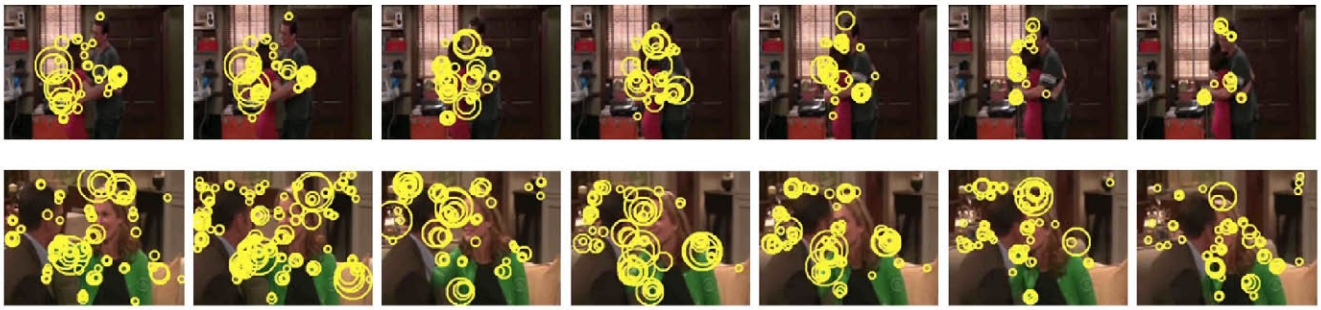

Figure 3. Detected STIP for two series of frames corresponding to the *hug* action (first row), and the *kiss* action (second row) from the TVHI database.

ensured by the Johnson-Lindenstrauss (JL) lemma [10]. Indeed, JL states that for any $0 < \varepsilon < 1$ and any integer N, there is a positive integer m such that

$$m \geq m_0 = \mathcal{O}(\frac{\ln N}{\varepsilon^2}). \tag{1}$$

Then, for any set $\mathcal{B}$ of N points in $\mathbb{R}^n$, there is a map $f : \mathbb{R}^n \to \mathbb{R}^m$, such that $\forall u, v \in \mathcal{B}$

$$\begin{aligned}(1-\varepsilon)\|u-v\|^2 \leq &\|f(u)-f(v)\|^2 \leq \\ &\leq (1+\varepsilon)\|u-v\|^2,\end{aligned} \tag{2}$$

and therefore a high-dimensional point cloud can be shrunk into a low-dimensional subspace, preserving the pair-wise distance between points. The variable ε measures the distortion introduced by the JL compression. Since a classifier depends only on the Euclidean distances between points to estimate the borders that separate the considered classes, both the original and the compressed domain are equivalent from a classifier point of view.

In particular, the linear map f can be represented by a matrix $\Phi \in \mathbb{R}^{m \times n}$. Moreover, it has been shown that every random matrix $\frac{1}{\sqrt{m}}\Phi$ verifies the JL lemma with high probability, as long as the probability distribution used to generate the elements of Φ satisfies some mild conditions (*i.e.* unit variance, zero mean, and bounded even moments) [1]. The Gaussian distribution is one example that fulfills the JL lemma, and the one used in our implementation to compute the matrix elements $\Phi_{i,j} \sim \mathcal{N}(0,1)$.

4. Recognition

The recognition task is carried out by the Citation-KNN classifier [27], which was introduced to adapt the K-NN classifier to the MIL framework. In particular, C-KNN is based on two main ideas regarding to K-NN. The first one consists of defining a new function that measures the distance between bags and is also robust to outliers. This distance is called *minimal Hausdorff distance*, whose mathematical expression is defined as:

$$\begin{aligned}h_1(A,B) &= \min_{a_i \in A} \min_{b_j \in B} \|a-b\| = \\ &= \min_{b_j \in B} \min_{a_i \in A} \|b-a\| = h_1(B,A),\end{aligned} \tag{3}$$

where A and B are two bags of instances: $A = \{a_1, ..., a_m\}$, and $B = \{b_1, ..., b_n\}$.

The second idea is to adopt the notion of *citation* from the bibliography field. Given an unlabeled bag b, this notion suggests to predict its label by considering not only the bags located in its nearest neighborhood (*references*), but also the bags that have b in their nearest neighborhood (*citers*). The number of positive and negative bags for the R-nearest references are defined as R_p and R_n respectively, and the number of positive and negative bags for the C-nearest citers are defined as C_p and C_n respectively. They are computed by the minimal Hausdorff distance. Finally, a positive label is assigned to bag b if the number of positive bags $(R_p + C_p)$ are larger than the number of negative bags $(R_n + C_n)$, and otherwise a negative label is assigned.

Since the default implementation of C-KNN is based on a binary classification [30], the *one-against-one* strategy is adopted in order to deal with the multi-class problem. This way, a voting strategy is applied over the predictions computed by the set of binary classifiers to make a final decision about the recognized human action in the video sequence.

5. Experimental results

In this section we evaluate the proposed human action recognition framework on the TVHI database [17], and compare it with other methods in the state of the art, also based on local spatio-temporal features. The TVHI database contains four types of human actions: *handshake*, *high five*, *hug*, and *kiss*. All the video sequences have been collected from 20 different TV shows, and therefore multiple challenges, such as different camera viewpoints, cluttered backgrounds, occlusions, and multiple moving people, appear in the scenes. Each class contains 50 video

sequences, which are further divided into two sub-sets: a training set containing the 80% of the sequences, and a test set containing the 20% of the sequences.

The *Average accuracy* metric is used to measure the recognition accuracy, which is defined as follows:

$$\text{Average accuracy} = \frac{\text{Total number of correct actions}}{\text{Total number of actions}} \quad (4)$$

Table 1 shows a comparison with other state-of-the-art approaches. The first method [28] combines STIP and HOG-HOF descriptors with BoW, however once the visual words have been computed, those which better represent the human action are selected. Another difference is that it uses an SVM classifier. The second approach [25] computes dense trajectories and describes them by means of the Motion Boundary Histogram (MBH) descriptor. It also adopts a BoW strategy together with an SVM classifier. On the other hand, two variations of the proposed system has been tested: one using a K-NN classifier (instead of the C-KNN one) and without applying the RP stage to reduce the feature vector dimension; and the other using a C-KNN classifier and without applying the RP stage. As can be observed, the proposed approach clearly outperforms all the other algorithms, reaching the best accuracy score, and improving the second best score by a 10.0%. This improvement can be attributed to the use of both the RP technique and the C-KNN classifier, as can be concluded from the accuracy scores obtained by the variations of the proposed system used in the comparison.

Regarding the RP approach, different values for the dimension (m) of the reduced feature vectors have been tested. Since feature vectors have a dimension of 500 before applying RP, a set of values ranging from 100 to 400 in steps of 100 have been evaluated. Table 2 shows the recognition accuracy obtained when using RP together with K-NN and C-KNN. As it can be seen, the recognition accuracy improves in some cases when using K-NN, however it is still very poor. In the case of C-KNN, the highest recognition accuracy is achieved when the reduced feature vectors have a dimension of 400.

m	**100**	**200**	**300**	**400**	**500 (no RP)**
K-NN	0.180	0.200	0.200	0.220	0.250
C-KNN	0.290	0.325	0.475	**0.675**	0.575

Table 2. Average accuracy obtained for different values of m in both classification schemes, K-NN and C-KNN.

Fig. 4 shows the empirical probability of the upper bound of the distortion suffered by a set of 100 points projected from $\mathbb{R}^{500} \rightarrow \mathbb{R}^m$, using a Gaussian projection matrix. For example, 95% of the points projected to $\mathbb{R}^{200}$ will suffer $\varepsilon = 0.2$ distortion or less.

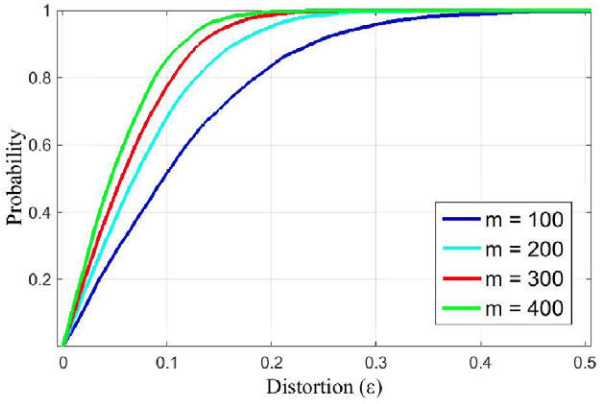

Figure 4. Distortion due to compression by RP.

6. Acknowledgements

This work has been partially supported by the Ministerio de Economía y Competitividad of the Spanish Government under the projects TEC2010-20412 (Enhanced 3DTV) and TEC2013-48453 (MR-UHDTV).

7. Conclusion

A new machine learning framework to recognize human actions is presented in this paper. It starts by detecting STIP and describing them by means of HOG-HOF descriptors. Then, a Bag-of-Words strategy is adopted to represent each video sequence by a compact feature vector. This is reduced in dimension by a random-projection technique that improves the performance. Finally, Citation K-NN algorithm is used for the classification of human actions, which is especially advantageous in MIL problems. Experimental results have shown that the presented action recognition methodology outperforms other state-of-the-art approaches, demonstrating its suitability for human-action recognition in realistic video sequences.

References

[1] R. I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *IEEE Symposium on Foundations of Computer Science*, pages 616–623, 1999.

[2] K. Avgerinakis, A. Briassouli, and I. Kompatsiaris. Recognition of activities of daily living. In *IEEE International Conference on Tools with Artificial Intelligence*, volume 2, pages 8–12, Nov 2012.

[3] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *International Conference on Knowledge Discovery and Data Mining*, pages 245–250, 2001.

[4] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta. Silhouette-based human action recognition using sequences

Feature representation	Classification strategy	Vocabulary	Accuracy
STIP + HOGHOF [28]	Filtered BoW + SVM	K=550	50.5%
Dense trajectories + MBH [25]	BoW + SVM	K=100	56.0%
Dense trajectories + MBH [25]	BoW + SVM	K=500	53.5%
STIP + HOGHOF	BoW + K-NN	K=500	25.0%
STIP + HOGHOF	BoW + C-KNN	K=500	57.5%
STIP + HOGHOF (our proposal)	BoW + RP + C-KNN	K=500	**67.5%**

Table 1. Average accuracy obtained with different human action recognition algorithms.

of key poses. *Pattern Recognition Letters*, 34(15):1799–1807, 2013.

[5] A. Ciptadi, M. S. Goodwin, and J. M. Rehg. Movement pattern histogram for action recognition and retrieval. In *Computer Vision*, pages 695–710. Springer, 2014.

[6] I. Everts, J. van Gemert, and T. Gevers. Evaluation of color stips for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2850–2857, Jun 2013.

[7] D. Geronimo and H. Kjellstrom. Unsupervised surveillance video retrieval based on human action and appearance. In *International Conference on Pattern Recognition*, pages 4630–4635, Aug 2014.

[8] N. Goel, G. Bebis, and A. Nefian. Face recognition experiments with random projection. In *Defense and Security*, pages 426–437, 2005.

[9] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo. Trajectory-based modeling of human actions with motion reference points. In *European Conference on Computer Vision*, pages 425–438. 2012.

[10] W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[11] I. Jolliffe. Principal component analysis. *Springer*, 1, 1986.

[12] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, pages 275–1, 2008.

[13] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Jun 2008.

[15] C. Li, B. Su, Y. Liu, H. Wang, and J. Wang. Human action recognition using spatio-temporal descriptor. In *International Congress on Image and Signal Processing*, volume 1, pages 107–111, Dec 2013.

[16] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *IEEE International Conference on Computer Vision*, pages 104–111, 2009.

[17] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. Structured learning of human interactions in tv shows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2441–2453, Dec 2012.

[18] A. Prest, V. Ferrari, and C. Schmid. Explicit modeling of human-object interactions in realistic videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):835–848, Apr 2013.

[19] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):601–614, Mar 2012.

[20] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *International conference on Multimedia Proceedings*, pages 357–360, 2007.

[21] K.-T. Song and W.-J. Chen. Human activity recognition using a mobile camera. In *International Conference on Ubiquitous Robots and Ambient Intelligence*, pages 3–8, Nov 2011.

[22] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2004–2011, 2009.

[23] C. Tomasi and T. Kanade. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[24] H. Uemura, S. Ishikawa, and K. Mikolajczyk. Feature tracking and motion compensation for action recognition. In *British Machine Vision Conference*, pages 1–10, 2008.

[25] H. W., A. Klaser, C. Schmid, and C.-L. L. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3169–3176, Jun 2011.

[26] S. Wan, M.-W. Mak, B. Zhang, Y. Wang, and S.-Y. Kung. Ensemble random projection for multi-label classification with application to protein subcellular localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5999–6003, 2014.

[27] J. Wang and J.-D. Zucker. Solving multiple-instance problem: A lazy learning approach. 2000.

[28] B. Zhang, F. De Natale, and N. Conci. Recognition of social interactions based on feature selection from visual codebooks. In *IEEE International Conference on Image Processing*, pages 3557–3561, Sep 2013.

[29] J.-T. Zhang, A.-C. Tsoi, and S.-L. Lo. Scale invariant feature transform flow trajectory approach with applications to human action recognition. In *International Joint Conference on Neural Networks*, pages 1197–1204, Jul 2014.

[30] Z.-H. Zhou and M.-L. Zhang. Ensembles of multi-instance learners. In *European Conference on Machine Learning*, pages 492–502. Springer, 2003.