Statistical Measures from Co-occurrence of Codewords for Action Recognition

Carlos Caetano, Jefersson A. dos Santos and William Robson Schwartz
Smart Surveillance Interest Group, Department of Computer Science,
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Keywords: Spatiotemporal Features, Co-occurrence, Action Recognition.

Abstract: In this paper, we propose a novel spatiotemporal feature representation based on co-occurrence matrices of codewords, called Co-occurrence of Codewords (CCW), to tackle human action recognition, a significant problem for many real-world applications, such as surveillance, video retrieval and health care. The method captures local relationships among the codewords (densely sampled), through the computation of a set of statistical measures known as Haralick textural features. We apply a classical visual recognition pipeline in which involves the extraction of spatiotemporal features and SVM classification. We investigate the proposed representation in three well-known and publicly available datasets for action recognition (KTH, UCF Sports and HMDB51) and show that it outperforms the results achieved by several widely employed spatiotemporal features available in the literature encoded by a Bag-of-Words model with a more compact representation.

1 INTRODUCTION

Human action recognition has been used in many real-world applications. In environments that require a higher level of security, surveillance systems can be used to detect and prevent abnormal or suspicious activities such as robberies and kidnappings. In addition, action recognition can be used in systems for video retrieval, this way an user could search for videos from the actions performed on it.

Considering only surveillance applications, such systems have traditionally relied on network cameras monitored by a human operator that must be aware of the actions carried out by people who are in the cameras field of view. With the recent growth in the number of cameras to be analyzed, the efficiency and accuracy of human operators has reached its limit (Keval, 2006). Therefore, security agencies have attempted computer vision-based solutions to replace or assist the human operator. In view of that, automatic recognition of suspicious activities is a problem that has attracted the attention of researchers in the area (Danafar and Gheissari, 2007; Xiang and Gong, 2008; Reddy et al., 2011; Wiliem et al., 2012).

Over the last decade, a significant portion of the progress in human action recognition has been achieved due to the design of novel discriminative local feature descriptors followed by a Bag-of-Words (BoW) model (Sivic and Zisserman, 2003) encoding. In general, such representations are based on spatiotemporal feature descriptors employed on the video domain (Krig, 2014).

The feature extraction process is very important since it is responsible for describing the video contents. These representations must be rich enough to allow proper recognition. To that end, local spatiotemporal features are the most popular descriptors for extracting information based on gradients or on motion analysis (Poppe, 2010). With such features extracted, usually, a codification is applied as an intermediate representation. The BoW model (Sivic and Zisserman, 2003) is the most common approach used to encode the features extracted from the videos. Such approach represents videos as histograms constructed from a set of visual features, known as visual dictionaries or codebooks. Finally, the intermediate representation is presented to a classification step to learn a function that can assign discrete labels to the videos.

Nowadays deep learning techniques, such as Convolutional neural networks (CNNs), have become a growing trend in computer vision. Such approaches are based on the employment of deeper and more complicated networks with a high number of parameters aiming to achieve higher accuracy. Nonetheless, most works forget that real world applications that involves recognition tasks such as robotics, self-

Caetano, C., Santos, J. and Schwartz, W.
Statistical Measures from Co-occurrence of Codewords for Action Recognition.
DOI: 10.5220/0006615803010308
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 5: VISAPP*, pages 301-308
ISBN: 978-989-758-290-5

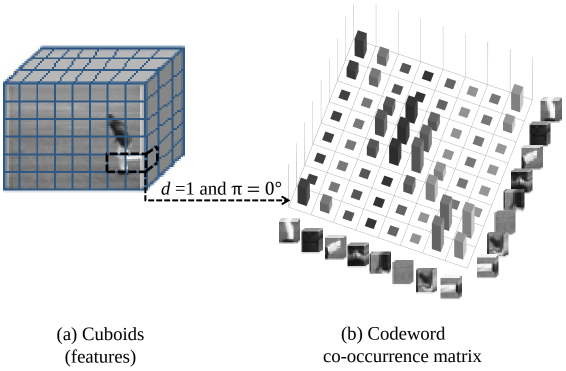

Figure 1: Creation a co-occurrence matrix from feature codewords. For this example, the comparison between pairs of cuboids is performed considering a 0° (horizontal) nearest neighbor with distance $d = 1$ (for more details, see Section 3). (a) Set of cuboids (features) densely extracted over the video. (b) Co-occurrence matrix computed from the feature codewords.

driving car and augmented reality, are usually executed on platforms with limited resources (Howard et al., 2017). Moreover, since the number of parameters is very large, they depend on heavy processing to converge. In addition, these networks need many labeled samples to learn, which is not always easily obtained in same applications. Because of these limitations, non-deep visual feature representation strategies are still needed.

Recently, a promising spatiotemporal local feature descriptor called Optical Flow Co-occurrence Matrices (OFCM) (Caetano et al., 2016) was developed. The method extracts a robust set of measures known as Haralick features to describe the flow patterns by measuring meaningful properties such as contrast, entropy and homogeneity of co-occurrence matrices to capture local space-time characteristics of the motion through the neighboring optical flow magnitude and orientation. OFCM has shown excellent results on the action recognition problem. Therefore, motivated by the great results achieved by OFCM, in this paper we propose a novel spatiotemporal feature representation called Co-occurrence of Codewords (CCW). The method is based on co-occurrence matrices derived from a precomputed codebook.

Our hypothesis for obtaining the CCW representation is based on the assumption that the information on a video sequence can be encoded by the spatial relationship contained on local neighborhoods of the features. More specifically, we assume that the information is adequately specified by a set codeword co-occurrence matrices computed for various angular relationships at a given offset between neighboring features. Therefore, matrices obtained by modifying the spatial relationship (different orientations or distances between features) will capture different information. Figure 1 illustrates the CCW representation computed based on the feature codewords provided by a pre-computed codebook for an horizontal displacement of a single cuboid (feature).

Given the aforementioned description of CCW, we consider three variations of the proposed representation. The first is based on the extraction of Haralick features (Haralick et al., 1973) from codeword co-occurrence matrices, the second employs a simple concatenation of the vectorized co-occurrence matrices, and the last combines the two previous variations.

According to the experimental results, the proposed feature representation followed by the SVM classifier is able to recognize actions accurately on KTH, UCF Sports and HMDB51 datasets. The employment of the CCW outperforms the results achieved by several widely employed spatiotemporal features available in the literature encoded by a BoW model (Sivic and Zisserman, 2003) and achieves comparable results with OFCM feature descriptor + BoW showing a more compact representation.

The remainder of this paper is organized as follows. In Section 2, we briefly review works in literature that explore co-occurrence to encode information. In Section 3, we introduce the novel spatiotemporal feature representation called Co-occurrence of Codewords (CCW). Then, Section 4 presents our experimental results, validating the high performance of the CCW feature representation. Finally, Section 5 presents conclusions obtained and future works.

2 RELATED WORK

In this section, we present a brief review of works that are close to the idea proposed in our approach by employing co-occurrence to encode information.

In (Zalevsky et al., 2005), the authors used co-occurrence for motion characterization. Their method allows the estimation of relative position, scale and the rotation of objects in the scene to obtain time to impact and relative distance between objects for navigation and collision avoidance.

The BoW model was improved in (Liu et al., 2008) by calculating spatial histograms where the co-occurrences of local features are considered to encode spatial relationship as 2^{nd} order features. Instead of assigning a local feature descriptor to a single codeword, one can assign it to the top-N closest codewords. The main difference between our work and (Liu et al., 2008) is that we consider the co-occurrence matrices of feature codewords while their final representation is composed assigning local features to *N* closest codewords.

A representation that characterizes photometric and geometric aspects of an image by computing co-occurrences of codewords with respect to spatial predicates over a hierarchical spatial partitioning of an image was proposed by (Yang and Newsam, 2011). Our proposed method differs from them since they combine different co-occurrence matrices with kernel functions and do not extract Haralick features.

A new image feature based on spatial co-occurrence among micropatterns was proposed in (Nosaka et al., 2012). To consider spatial relations, the authors extract co-occurrence matrices among multiple micropatterns, represented by Local Binary Pattern (LBP), and applied it to face recognition and texture recognition tasks. Although their method is similar to our approach, they do not perform co-occurrence on codewords and also do not extract Haralick features from the matrices.

In (Zhang et al., 2014), the co-occurrence was applied to the person re-identification problem. The author creates a codeword image in which each pixel represents the centroid of a patch that has been mapped to a codeword. The appearance transformation between camera views is encoded by a co-occurrence matrix computed by the inner product of the codeword probe and gallery images in the reproducing kernel Hilbert space (RKHS). Later, the authors extended their model to multi-shot scenarios (Zhang and Saligrama, 2017).

The use of co-occurrence was also employed in action recognition tasks. In (Banerjee and Nevatia, 2011), a Minimum Spanning Tree (MST) based distance function was used to count the co-occurrence of Spatio Temporal Interest Points (STIP) (Laptev, 2005) codewords in terms of the edge connectivity of latent variables of a Conditional Random Field (CRF) classifier. The latent semantic analysis (LSA) was applied in (Zhang et al., 2012) to exploit high order co-occurrence by mapping codewords into a co-occurrence space. In (Sun and Liu, 2013), the authors model the semantic relationship (spatial and temporal) of codewords in terms of normalized googlelike distance (NGLD), which measures the co-occurrence frequency of each pairwise codewords in the video. A new representation called directional pairwise feature (DPF) was proposed by (Liu et al., 2014) which uses direction instead of distance when describing the pairwise co-occurrence and also pairwise counts.

Even though the aforementioned methods used co-occurrence to encode information, they do not compute co-occurrence matrices considering different angles (more details in Section 4) on their description process. Moreover, as the main difference between our work and others is that we consider the co-occurrence of feature codewords and compute the Haralick features from the co-occurrence matrices as the final feature vector.

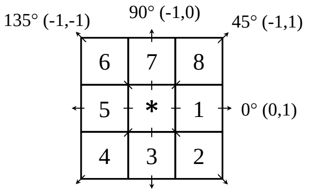

Figure 2: Offset configurations with $d = 1$. Cells 1 and 5 are the 0° (horizontal) nearest neighbors to cell *; cells 2 and 6 are the 135° nearest neighbors; cells 3 and 7 are the 90° nearest neighbors; and cells 4 and 8 are the 45° nearest neighbors to *. Note this information is purely spatial, and has nothing to do with pixel intensity values (Haralick et al., 1973) (Figure from (Caetano et al., 2016)).

3 PROPOSED APPROACH

Several action recognition approaches, such as (Wang et al., 2011) and (Wang et al., 2009), are based on feature extraction followed by a BoW representation to encode the information extracted from the video. While those works have demonstrated encouraging recognition accuracy, a rich source of information contained in these methods, such as spatial relations contained on the features, have not been fully explored.

To explore the local spatial relations contained of the features, we propose a novel spatiotemporal feature representation, called Co-occurrence of Codewords (CCW), based on the co-occurrence matrices computed over feature codewords. Such co-occurrence matrices express the distribution of the features at a given offset over feature codewords from a pre-computed codebook, as illustrated in Figure 1.

Our hypothesis to design the CCW representation is based on the assumption that the information on a video sequence can be encoded by the overall relationship of the feature codewords from a pre-computed codebook. In addition, we believe that it can be specified by a set of codeword dependence matrices computed for various angular relationships and distances between neighboring codewords on the video. Once the co-occurrence matrices have been computed, we extract a set of measures known as Haralick textural features (Haralick et al., 1973) to describe the patterns.

The classical textural feature was proposed by Haralick et al. (Haralick et al., 1973) and is based on the gray level co-occurrence matrix (GLCM) which estimates the joint distribution of pixel intensity given

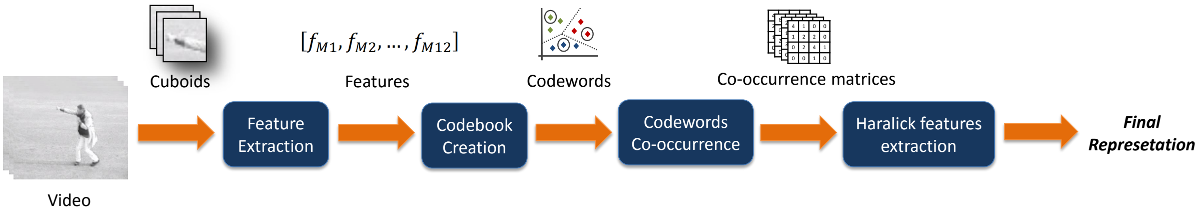

Figure 3: Diagram illustrating the pipeline extraction of the proposed spatiotemporal feature representation.

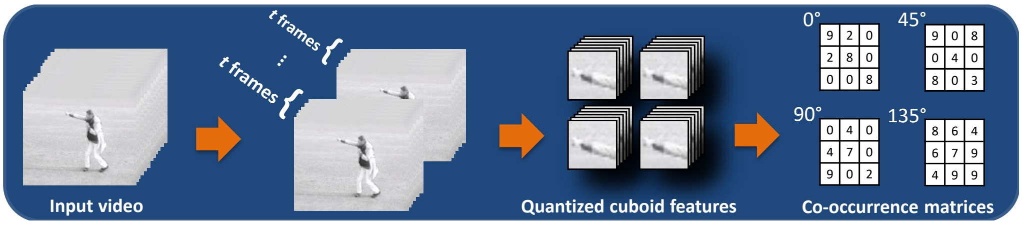

Figure 4: Detailed “Codewords Co-occurrence” illustration.

a distance and an orientation. Mathematically, a co-occurrence matrix Σ is defined over an $n \times m$ image I, at a specified offset (Δ_x, Δ_y), as

$$\Sigma_{\Delta_x,\Delta_y}(i,j)=\sum_{r=1}^{n}\sum_{q=1}^{m}\begin{cases}1, & \text{if } I(r,q)=i \text{ and}\\ & I(r+\Delta_x,q+\Delta_y)=j,\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

where i and j are the image intensity values separated by distance d, r and q are the spatial positions in the image I and the offset (Δ_x, Δ_y) depends on the angle π used. Usually, π is expressed as angles 0° $(0,d)$, 45° $(-d,d)$, 90° $(-d,0)$ and 135° $(-d,-d)$. Figure 2 illustrates possible offset configurations. Note that in the proposed method, we do not compute the matrices using the image intensity values, but using quantized features (codewords), as it will be discussed in the next paragraphs.

Aiming to depict the meaningful properties contained in the co-occurrence matrices, Haralick et al. introduced 14 statistical measures that can be extracted from the computed matrices: angular second moment, contrast, correlation, sum of squares, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, two information measures of correlation, and maximal correlation coefficient.

The process of computing the CCW representation is illustrated in Figure 3 and will be explained in details as follows. First, a dense sampling step is applied to the video dividing it into $n_i \times n_j \times n_t$ regions. These regions are referred to as cuboids and are described by their width (n_i), height (n_j), and length (n_t). With the cuboids at hand, we apply a spatiotemporal feature descriptor in order to describe the information of each cuboid.

Since the features obtained from the cuboids are composed by real valued vectors, a quantization step is applied to compute the co-occurrence matrices. The feature quantization used is based on a codebook, or visual dictionary. Let $\mathcal{C}$ be a visual codebook obtained by an unsupervised learning algorithm (e.g., k-means clustering algorithm). $\mathcal{C} = \{\mathbf{c}_k\}$, $k \in \{1,\dots,K\}$, where $\mathbf{c}_k \in \mathbb{R}^D$ is a codeword and K is the number of codewords.

Given a video, we compute α co-occurrence matrices for each t frames from the quantized features according to Equation 1. Here, α is the number of angles used ($\pi = 0^\circ$, 45°, 90° and 135°) and t is according to the temporal length of the cuboids. Then, the co-occurrence matrices are accumulated according to its angles. After that, we extract f Haralick textural features (Haralick et al., 1973) for each co-occurrence matrix, generating a feature vector with f dimensions per matrix, where f is the number of extracted Haralick features. Finally, all feature vectors are concatenated in $X = \{x_i\}, i \in \{1,...N\}$, followed by z score norm as

$$z_i = \frac{x_i - \mu_i}{\sigma_i}, \tag{2}$$

where $x_i \in \mathbb{R}^D$ represents each dimension of the concatenated vectors and N is the length of $\alpha \times f$, μ_i is the mean value of the dimension i and σ_i is standard deviation of each dimension i. Such process is the “Codewords Co-occurrence” block shown on Figure 3 and illustrated with more details in Figure 4.

We also present two variations of the CCW representation: (i) a simply concatenation of the vectorized co-occurrence matrices with no Haralick feature extraction (inspired by the work (Nosaka et al., 2012)); and (ii) a combination of the Haralick features extraction and the concatenation of the vectorized co-occurrence matrices. The vectorized matrices can be seen as an histogram in which each bin is composed by a pair of features instead of just a single feature.

4 EXPERIMENTS

In this section, we describe the datasets used for the evaluation as well as the evaluation protocol. We evaluate the proposed representation and compare it to a Bag-of-Words based action recognition approach and to methods of the literature that also employ co-occurrence to encode information.

4.1 Action Recognition Datasets

KTH is a well-known and publicly available dataset for action recognition (Schuldt et al., 2004). It consists of six human action classes: walking, jogging, running, boxing, waving and clapping. Each action is performed several times by 25 subjects in four different scenarios. In total, the data consist of 600 videos and spatial resolution of 160×120 pixels. To obtain fair comparison, we follow the experimental setup used by Wang et al. (Wang et al., 2009) and divide the samples into training/validation set $(8+8$ people) and test set (9 people). The performance is evaluated as suggested in (Schuldt et al., 2004), i.e., by reporting the average accuracy over all classes.

UCF Sports is a realistic dataset (Rodriguez et al., 2008) that consists of a set of actions collected from various sports. It is composed by ten different types of human actions: swinging (on the pommel horse and on the floor), diving, kicking (a ball), weight-lifting, horse-riding, running, skateboarding, swinging (at the high bar), golf swinging and walking. It consists of 150 video samples with a frame rate of 10 fps and spatial resolution of 720×480 pixels. We use a leave-one-out setup as suggested by the authors (Rodriguez et al., 2008). The performance metric used is the average accuracy over all classes.

HMDB51 (Kuehne et al., 2011) is also a realistic and challenging activity dataset composed of video clips from movies, the Prelinger archive, Internet, Youtube and Google videos, and comprised of 51 activity categories. It consists of $6,766$ activity samples with a resolution of 240 pixels in height with preserved aspect ratio. We follow the original protocol using three train-test splits. The performance is evaluated by computing the average accuracy across all classes over the three splits.

4.2 Experimental Setup

In the interest of a fair comparison, we apply the same evaluation pipeline as (Wang et al., 2009). It is a classical visual recognition pipeline in which involves two phases: training and testing.

In the training phase, we first densely extract OFCM spatiotemporal feature descriptors (Caetano et al., 2016)[1]. Dense sampling extracts video blocks at regular positions in space and time. There are 3 dimensions to sample from: $n_i \times n_j \times n_t$. In our experiments, the minimum size of a block is 18×18 pixels and 10 frames. Spatial and temporal sampling are performed with 50% of overlapping. Next, following the visual recognition strategy, the local features must be encoded into a mid-level representation to be used for the classification task. However, a visual codebook must be created before the encoding. Thus, we randomly sample K training features. This is very fast and according to (Kläser et al., 2008) the results are very close to those obtained using vocabularies built with k-means. After, for each video sequence, we extract a Co-occurrence of Codewords (CCW) representation. Spatiotemporal features are first quantized into the codewords according to the codebook previously created and a video is then represented as 4 co-occurrence matrices ($\pi = 0°$, $45°$, $90°$ and $135°$). Here, we extract $f = 12$ Haralick textural features: angular second moment, contrast, correlation, sum of squares, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy and maximal correlation coefficient. Then, the feature vectors are assigned to their closest codeword using Euclidean distance. Finally, one-against-all classification is performed by a non-linear Support Vector Machines (SVM) with a RBF-kernel.

In the testing phase, a new test video sequence is classified by applying the trained classifier obtained during the training phase. Thus, for the test video sequence, OFCM spatiotemporal feature descriptors are extracted with dense sampling. Next, the CCW representation is generated using the visual codebook previously created. Then, that feature vector is given as input to the trained classifier to predict the class label of the test video sequence.

It is important to emphasize that in the experiments we change only the mid-level representation used in the pipeline since our main goal is to compare the real contribution of our proposed feature representation, i.e., for every experiment, the pipeline is the same only the feature representation is changed.

[1] Although we used OFCM, it is important to emphasize that any feature descriptors can be used with CCW.

Table 1: Video classification Acc. (%) results of the combined variations of the CCW representation on KTH actions dataset (Schuldt et al., 2004).

	Raw concatenation	L_1 **norm**	L_2 **norm**	z **score norm**
Approach	**Acc. (%)**	**Acc. (%)**	**Acc. (%)**	**Acc. (%)**
Vectorized + CCW + OFCM ($k = 4$)	72.22	72.22	72.22	91.20
Vectorized + CCW + OFCM ($k = 8$)	88.58	78.55	78.55	93.06
Vectorized + CCW + OFCM ($k = 18$)	88.43	81.02	81.02	**96.30**
Vectorized + CCW + OFCM ($k = 32$)	87.50	78.55	78.55	94.75

Table 2: A summary of the proposed feature vector lengths of the CCW representation and OFCM + BoW model fine tuned on KTH actions dataset (Schuldt et al., 2004).

	Approach	**Codewords** k	**Feature length**
Published results	BoW + OFCM (Caetano et al., 2016)	4000	4000
Our results	CCW + OFCM (Haralick feature extraction)	**28**	**48**
	CCW + OFCM (Vectorized matrices)	16	1024
	CCW + OFCM (Concatenation + z score norm)	18	1344

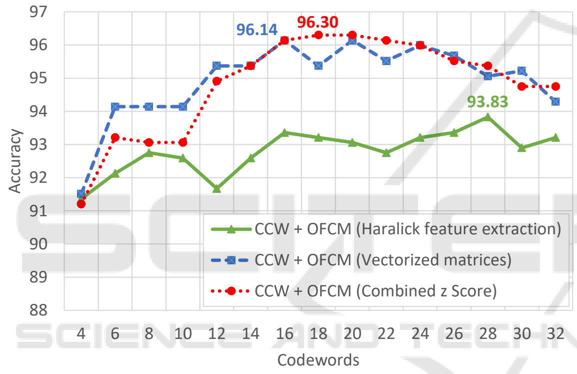

Figure 5: Accuracy of codebook size variation on KTH dataset (Schuldt et al., 2004).

4.3 Results and Comparisons

In this section, we present experiments for parameters optimization and report a comparison of our proposed feature representation. We used the KTH dataset to perform parameter setting and then used such parameters on UCF Sports and HMDB51 dataset experiments. We focused on the optimization of the number of codewords K used to quantize the features and also to create the co-occurrence matrices. We used the optimized parameters of the OFCM feature descriptor for KTH presented in (Caetano et al., 2016).

Figure 5 shows the codewords K variation for three variations of our proposed feature representation. For the first one, CCW + OFCM (Haralick feature extraction), $K = 28$ presents the best accuracy value reaching 93.83% with final feature vector length of 48 (4 co-occurrence matrices × 12 Haralick features). Secondly, CCW + OFCM (vectorized matrices), reaches an accuracy of 96.14% with $K = 16$ and a final feature vector length of 1024 (4 co-occurrence matrices × 16 × 16).

We also present experiments with a third variation of the CCW representation in which is a combination of the two aforementioned variations. Here, we simply concatenated the last two presented approaches (Haralick and Vectorized). For that purpose, we empirically tested four different combination strategies: a raw concatenation; concatenation followed by L_1 norm; concatenation followed by L_2 norm; and concatenation followed by z score norm. For the z score norm, we learn the means μ_i and standard deviations σ_i of each feature dimension (x_i) during the training phase.

Table 1 shows the results of the four combination strategies with the best result of 96.30% being achieved by the combined CCW representation followed by z score norm with $K = 18$. Figure 5 also illustrates the codewords K variation for CCW + OFCM (Combined z score). Our method presents a more compact representation, using just $k = 18$ codewords and a final feature vector length of 1344 (4 co-occurrence matrices × (18 × 18) + 48 Haralick features) when directly compared to the BoW models used in the literature.

Table 2 summarizes the feature vector lengths for the proposed CCW and for the BoW representation length used in the literature (Caetano et al., 2016). We can see that CCW presents a more compact representation using fewer codewords and with a smaller final feature vector length. For instance, CCW + OFCM (Combined z score) achieves the same accuracy as the BoW + OFCM using a more compact representation with just $k = 18$ codewords and a final feature vec-

Table 3: Action recognition accuracy (%) results of OFCM + CCW representation and classic spatiotemporal features of the literature + BoW on KTH (Schuldt et al., 2004) and UCF Sports (Rodriguez et al., 2008) action datasets. Results for HOG, HOF, HOG/HOF and HOG3D were obtained from (Wang et al., 2009).

		KTH	UCF Sports	HMDB51
	Approach	**Acc. (%)**	**Acc. (%)**	**Acc. (%)**
Published BoW results	BoW + HOG (Dalal and Triggs, 2005)	79.00	77.40	28.40
	BoW + HOF (Laptev et al., 2008)	88.00	84.00	35.50
	BoW + HOG/HOF (Laptev et al., 2008)	86.10	81.60	43.60
	BoW + HOG3D (Kläser et al., 2008)	85.30	85.60	36.20
	BoW + MBH (Dalal et al., 2006)	89.04	90.53	51.50
	BoW + GBH (Shi et al., 2015) [1]	92.70	-	38.80
	BoW + OFCM (Caetano et al., 2016)	**96.30**	**92.80**	**56.91**
	Dense Trajectories (Wang et al., 2011)	94.20	88.20	46.60
Published co-occurrence results	STIP + CRF (Banerjee and Nevatia, 2011)	93.98	-	-
	CS + HOG/HOF (Zhang et al., 2012)	91.20	-	26.82
	ST-NGLDC + 3D-SIFT (Sun and Liu, 2013)	91.82	89.74	-
Our results	CCW + OFCM (Haralick feature extraction)	93.83	90.53	52.62
	CCW + OFCM (Vectorized matrices)	96.14	**91.33**	55.08
	CCW + OFCM (Combined)	**96.30**	91.07	**55.25**

tor length of 1344, instead of $k = 4000$ used by the OFCM + BoW.

We also compare our approach with several classic local spatiotemporal features + BoW model of the literature. According to Table 3, CCW + OFCM (Combined) achieved the higher accuracy value, which is the same as the BoW + OFCM, reaching 96.30% on KTH dataset. Moreover, we note an improvement of 2.10 percentage points (p.p.) on the KTH dataset, achieved by our representation when compared to Dense Trajectories method (Wang et al., 2011). Furthermore, it is important to point out that their approach uses a combination of three different feature descriptors (HOG, HOF and MBH), while we only used the OFCM feature. Although our CCW representation did not achieve better results on UCF Sports and HMDB51 datasets, it still presents comparable results being only 1.47 p.p. behind the best result on UCF Sports and 1.66 p.p. on HMDB51. Furthermore, as already mentioned it is worth emphasizing that the CCW has a more compact representation (smaller feature vector length) as shown on Table 2.

A comparison with methods of the literature that also employed co-occurrence to encode information on action recognition tasks is shown on the second part of Table 3. On the KTH dataset, our CCW representation (Vectorized and Combined) achieves an improvement of 2.32 p.p. when compared to the STIP + CRF (Banerjee and Nevatia, 2011). Moreover, on UCF Sports, we outperform Spatio Temporal Nor- malized Google-Like Distance Correlogram (ST-NGLDC) method (Zhang et al., 2012) by 2.93 p.p. using our CCW (Vectorized) representation. Finally, our three CCW evaluated representations presented a large accuracy gain when compared to Co-occurrence Space (CS) method (Zhang et al., 2012).

The improvement achieved regarding the compactness of CCW representation over BoW model is very significant. We believe the reason for such improvement lies on the information extracted from co-occurrence related to global structures in various local region-based features, moreover a pair (co-occurrence) of codewords have more "vocabulary" than a single histogram bin of codewords. In this way, the CCW representation can encode the information in more details than the BoW model which is a histogram based features that use single codewords discarding important information concerning spatial relations among the features.

5 CONCLUSIONS

We have proposed a novel spatiotemporal feature representation called Co-occurrence of Codewords (CCW). The method is based on the extraction of Haralick features from the co-occurrence matrices derived from neighboring cuboids (features) obtained from a pre-computed codebook.

We have demonstrated that CCW representation is able to recognize actions accurately on three well-known datasets. The employment of the CCW outperforms the results achieved by several widely employed spatiotemporal features available in the literature encoded by a Bag-of-Words (BoW) model and

[2] The results presented in (Shi et al., 2015) are using FV. However, since we directly compare with BoW, here we apply the same mid-level representation (BoW).

achieves comparable results with OFCM feature descriptor + BoW showing a more compact representation. In addition, CCW presented better recognition accuracy when compared to methods of the literature that also employed co-occurrence to encode information on action recognition tasks

Possible directions for future works include to evaluate other features with CCW representation. Moreover, we would like to evaluate CCW in other video-related tasks. It is important to emphasize that, since the CCWis a spatiotemporal feature representation, it can be also applied to other computer vision applications involving video description.

ACKNOWLEDGMENTS

The authors would like to thank the Brazilian National Research Council – CNPq (Grants #311053/2016-5 and #449638/2014-6), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17) and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project).

REFERENCES

Banerjee, P. and Nevatia, R. (2011). Learning neighborhood cooccurrence statistics of sparse features for human activity recognition. In *AVSS*.

Caetano, C., dos Santos, J. A., and Schwartz, W. R. (2016). Optical flow co-occurrence matrices: A novel spatiotemporal feature descriptor. In *ICPR*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*.

Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *ECCV*.

Danafar, S. and Gheissari, N. (2007). Action recognition for surveillance applications using optic flow and svm. In *ACCV*.

Haralick, R. M., Shanmugam, K. S., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*.

Keval, H. (2006). Cctv control room collaboration and communication: Does it work? In *Human Centred Technology Workshop*.

Kläser, A., Marszałek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *BMVC*.

Krig, S. (2014). Interest point detector and feature descriptor survey. In *Computer Vision Metrics*. Apress.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). Hmdb: A large video database for human motion recognition. In *ICCV*.

Laptev, I. (2005). On space-time interest points. *IJCV*.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *CVPR*.

Liu, D., Hua, G., Viola, P., and Chen, T. (2008). Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*.

Liu, H., Liu, M., and Sun, Q. (2014). Learning directional co-occurrence for human action classification. In *ICASSP*.

Nosaka, R., Ohkawa, Y., and Fukui, K. (2012). Feature extraction based on co-occurrence of adjacent local binary patterns. In *PSIVT*.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image Vision Comput.*

Reddy, V., Sanderson, C., and Lovell, B. (2011). Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In *CVPRW*.

Rodriguez, M., Ahmed, J., and Shah, M. (2008). Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*.

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: A local svm approach. In *ICPR*.

Shi, F., Laganiere, R., and Petriu, E. (2015). Gradient boundary histograms for action recognition. In *WACV*.

Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *ICCV*.

Sun, Q. and Liu, H. (2013). Learning spatio-temporal co-occurrence correlograms for efficient human action classification. In *ICIP*.

Wang, H., Klaser, A., Schmid, C., and Liu, C.-L. (2011). Action recognition by dense trajectories. In *CVPR*.

Wang, H., Ullah, M. M., Klaser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *BMVC*.

Wiliem, A., Madasu, V., Boles, W., and Yarlagadda, P. (2012). A suspicious behaviour detection using a context space model for smart surveillance systems. *CVIU*.

Xiang, T. and Gong, S. (2008). Video behavior profiling for anomaly detection. *TPAMI*.

Yang, Y. and Newsam, S. (2011). Spatial pyramid co-occurrence for image classification. In *ICCV*.

Zalevsky, Z., Rivlin, E., and Rudzsky, M. (2005). Motion characterization from co-occurrence vector descriptor. *PRL*.

Zhang, L., Zhen, X., and Shao, L. (2012). High order co-occurrence of visualwords for action recognition. In *ICIP*.

Zhang, Z., Chen, Y., and Saligrama, V. (2014). A novel visual word co-occurrence model for person re-identification. In *ECCVW*.

Zhang, Z. and Saligrama, V. (2017). Prism: Person reidentification via structured matching. *TCSVT*.