Combining Video Subsequences for Human Action Recognition

Leonardo Onofri, Paolo Soda

Computer Science & Bioinformatics Laboratory
Università Campus Bio-medico di Roma, Integrated Research Centre
{l.onofri,p.soda,g.iannello}@unicampus.it

Abstract

Human action recognition is an active area with applications in several domains such as visual surveillance, video retrieval and human-computer interaction. Current approaches assign action labels to video streams considering the whole video as a single image sequence. Such approaches, albeit very refined, may fail on some samples due to large variability between frames, suggesting that features extracted from training videos may adopt description and classification models that may better represent video portions rather than the whole stream. To this aim, in this paper we propose a multiple subsequence combination method dividing the video into a number of consecutive subsequences and classifying each one applying a part-based method in conjunction with the bag of visual words approach. We classify a video combining subsequence labels according to rules inspired by the multiple expert system framework. We extensively tested our approach on the KTH, UCF sport and Youtube datasets showing, on the one side, that it outperforms a method classifying action using the whole stream and, on the other side, that its performance are robust and stable throughout all the datasets since our best results are comparable with the best published ones.

1 Introduction

Human action recognition is an active research area with many related topics, such as human detection and tracking, person identification and human-computer interaction, to name a few. Existing approaches can be roughly divided into part-based and holistic methods, and the results reported in the literature show that the former outperform the latter [1, 2, 3, 8, 10, 12, 15, 18]. Part-based approaches consider a space-time video volume as a collection of local parts, where each one consists of some distinctive motion patterns. Each part is represented by local descriptors, which are then quantized into a vocabulary composed of visual words, according to the bag of visual words (BOVW) approach [21]. A video is then represented as a vector containing the visual word occurrence, which is used as feature vector in the classification task. Existing works apply this procedure extracting the local features from the entire video sequence but, in several cases, the video large variability could lead to a decrease in performance. Furthermore, such approaches cannot be applied when the human action recognition system can tolerate a limited delay. For instance, in a video surveillance scenarios, the operator must be quickly alerted in case of any dangerous situations.

The main contribution of this paper is the definition of a method for multiple subsequence combination (MSC). Its rationale comes from observing that performing human action recognition on a single video subsequence composed of $n$ frames provides satisfactory performance [5, 11, 14]. Therefore, when a video longer than $n$ frames is available, we can analyze multiple subsequences to improve the recognition performance. Indeed, on the one hand features extracted from training samples may adopt description and classification models which may better represent video portions rather than the whole stream, on the other hand the fusion of decisions on multiple subsequences may further increase recognition performance, similarly to what has already been proven in the multiple expert system (MES) framework [6, 7]. The MSC approach classifies human action in a video combining the degree of support to the hypothesis that in each video subsequence one action is performed, applying rules inspired by MES. In this way the robustness of the classification increases as the stream is collected, allowing a trade-off between accuracy and time constrains.

As a second contribution, to the best of our knowledge in this paper we report the first attempt integrating two approaches for: (i) local part description, computing SIFT descriptors extended to spatio-temporal do-
main [2], (ii) vocabulary size reduction through the information bottleneck (IB) method [16], which overcomes BOVW limitations [10].

We extensively tested our MSC approach on the three most used action datasets, i.e. KTH [15], UCF sport [13] and Youtube [9], achieving three main results. First, we verify that classifying human action in a video through the combination of several subsequences improves the recognition performance in comparison to classifying all video frames at the same time. Second, we show that the application of the MSC approach using a small number of consecutive subsequences provides a recognition accuracy comparable to the one achieved classifying the video as a whole. This result may be exploited in applications where time constraints play a fundamental role. Third, our best results are comparable with the best published ones for each dataset.

2 MSC Approach

In this section we present the approach we use to combine subsequences of the video to get the final classification of an action. We omitted details on local part description, BOVW and IB methods. The interested reader may refer to [2], [21] and [10], respectively.

The rationale of our approach comes from observing that a video may show large variability and, hence, in many cases a system recognizing a single sequence, albeit very refined, may fail to achieve high performance. Indeed, the recognition system, on the basis of the features extracted from training samples, adopts description and classification models which, as a whole, may be not adequate for all frames, while they may better represent its portions. The idea of combining various classification systems aiming at compensating for the weakness of one expert while preserving its own strength has been widely investigated in MES framework [6, 7]. Here, we regard the classification of several subsequences as a multiple classification task, where we combine the labels assigned to each subsequence of the video. Indeed, a suitable combination of the outputs assigned to a set of subsequences could provide better performance than those obtained labeling all video frames as a whole.

Let us introduce the following notation:

- \( v \) is an \( N \) frames long video;
- \( s_i \) is the \( i \)-th subsequence extracted from the video \( v \), with \( i = 1, 2, \ldots, L \), \( L \leq \lfloor N/n \rfloor \), \( n \) is the number of frame in \( s_i \), and \( s_j \cap s_i = \emptyset \) for \( j \neq i \);
- \( x_i \) is the feature vector computed from \( s_i \);
- \( \Omega = \{ \omega_k \} \) is the set of class labels, with \( k = 1, 2, \ldots, c \), where \( c \) is the number of classes;
- \( VP(v) \) is the video profile of video \( v \), defined as follows:

\[
VP(v) = \begin{bmatrix}
d_1(x_1) & \ldots & d_{\ell}(x_1) \\
\vdots & \ddots & \vdots \\
d_1(x_L) & \ldots & d_{\ell}(x_L)
\end{bmatrix}
\]  

(1)

where the element \( d_j(x_i) \) is any measure in \([0, 1]\) representing the support to the hypothesis that subsequence \( s_i \) of the video \( v \) belongs to the class \( j \).

On this basis, the entries of \( VP(v) \) can be combined according to different rules to get the final classification of \( v \). In this respect, we apply the following five popular criteria inspired by the MES framework [6, 7]: average rule, maximum rule, minimum rule and product rule. Each one defines a criterion to get a vector \( M(v) \) whose elements

\[
\mu_k(v) = \mathcal{F}(d_k(x_1), \ldots, d_k(x_L)), \quad k = 1, \ldots, c, \tag{2}
\]

represent the degree of support for each class. The crisp label assigned to \( v \) by the combination rule is \( \omega_s \) if \( \mu_s(v) \geq \mu_k(v) \), \( \forall k = 1, \ldots, c \), where ties are resolved arbitrarily.

The average, maximum, minimum and product rules compute the corresponding elements of \( M(v) \), denoted as \( \mu_k^{\text{avr}}(v) \), \( \mu_k^{\text{max}}(v) \), \( \mu_k^{\text{min}}(v) \) and \( \mu_k^{\text{prod}}(v) \), respectively, combining each column of \( VP(v) \) as follows:

\[
\mu_k^{\text{avr}}(v) = \frac{1}{L} \sum_{i=1}^{L} d_k(x_i), \tag{3}
\]

\[
\mu_k^{\text{max}}(v) = \max \{d_k(x_1), d_k(x_2), \ldots, d_k(x_L)\}, \tag{4}
\]

\[
\mu_k^{\text{min}}(v) = \min \{d_k(x_1), d_k(x_2), \ldots, d_k(x_L)\}, \tag{5}
\]

\[
\mu_k^{\text{prod}}(v) = \prod_{i=1}^{L} d_k(x_i). \tag{6}
\]

3 Datasets

MSC approach has been tested on the three most used action datasets, namely KTH, UCF sport and Youtube.

The KTH dataset [15] consists of six human action classes: walking, jogging, running, boxing, waving and clapping. Each action is performed by 25 subjects in
four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. We adopt the leave-one-person-out cross validation, as recommended in [4].

The UCF sport dataset [13] contains ten 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. We adopt a leave-one-out cross validation.

The Youtube dataset [9] consists of eleven human action classes: basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. Each action videos are grouped in 25 different scenarios sharing some common features. We follow the original setup using a leave-one-group-out cross validation.

4 Results

In this section we briefly describe the explored MSC parameter configurations, and then we report the results.

We carried out tests exploring different settings of our approach. With respect to vocabulary size, we test different setups where the number of initial and final codebook words range in $[1500, 3000]$ and $[400, 1000]$, respectively. We note that every dataset shows a fairly strong dependence on the size of the vocabulary, confirming what was reported in previous works [2, 3, 8, 10, 12, 15, 18]. Furthermore, codebook size reduction improves the performance of 8.0%, on the average. We also evaluate the influence of subsequence length ($n$) and number of subsequences used ($L$). We varied $n$ in $[20, 150]$ and $L$ in $[1, \lfloor N/n \rfloor]$, where $N$ is the length of the whole video. In this regard, Figure 1 shows the influence of different values of $L$ when we use the most performing value of $n$ on each dataset. Finally, we evaluate the four combination rules presented in previous section, i.e. average rule, maximum rule, minimum rule and product rule.

Among the results provided by each configuration, in the following we present the best ones only. We observe that they have been achieved when $L = \lfloor N/n \rfloor$ and using the product combination rule.

We discuss the results with respect to the three main contributions claimed in section 1. The first issue claims that classifying human action in a video through the combination of several subsequences improves the recognition performance in comparison to classifying all video frames at the same time. In this respect, Table 1 shows that using all the subsequences extracted from the video, our MSC approach achieves a recognition rate equal to 97.0%, 88.0% and 84.4% on the KTH, UCF sport and Youtube dataset, respectively. Such results decrease to 95.8%, 86.7% and 75.3% if we apply the same BOVW approach to classify all the video frames at the same time. As the most important result of our contribution, we observe that MSC approach always outperforms the approach which classifies each video as a whole. Such an improvement is particularly evident in the Youtube dataset because the greater video length permits to get the most benefit from dividing a video into subsequences.

Let us now turn our attention on the application of the MSC approach using a small number of consecutive subsequences. Note that this issue is important in application domains whose constrains ask for minimizing the number of frames to be processed to recognize an action, as remarked in section 1. When the number of processed frames decreases, video descriptors may catch less information, and the performance of classification methods may significantly drop off. In this respect, most of existing works process offline the whole video sequences, classifying human action using all the frames composing the videos [1, 2, 10, 12, 18, 19]. One major advantage of our MSC algorithm is that it permits to use a variable number of subsequences with variable length, allowing the user to set a trade-off between recognition accuracy, domain constrains and computational load. For instance, classifying KTH dataset using three subsequences composed of 60 frames provides a recognition rate equal to 96.5%, on the Youtube dataset we achieve a recognition rate equal to 75.9% using four subsequences composed of 150 frames, while on UCF dataset we obtain a recognition rate equal to 87.3% using three subsequence composed of 35 frames. It is worth noting that such performance are always larger.
than those achieved classifying the video as a whole (fourth line of Table 1).

Finally, with respect to the third claim of section 1, i.e. that MSC best performance are comparable with the state-of-the-art, let us refer again to Table 1 (lines 2, 3 and 5). Comparing MSC recognition rates with the best ones reported in the literature on the three datasets [17, 20] we observe that its performance are robust and stable since its results are close to the best published ones and it outperforms the recognition rates achieved by [17] on UCF sport and by [20] on KTH and Youtube.

5 Conclusions

Existing approaches for human action recognition assign labels to video streams considering the whole video as a single image sequence. To cope with variability between frames we proposed a MES-like method combining multiple consecutive video subsequences, which adopts a part-based approach for subsequence description, and a vocabulary size reduction algorithm for BOVW classification. The results show that the proposed approach: (i) provides larger recognition performance than those obtained classifying the video as a whole, (ii) when applied using a small number of consecutive subsequences provides a recognition accuracy comparable to the one achieved classifying the video as a whole, and (iii) provides performance comparable with the best published ones for each dataset.

Future works are directed towards the analysis of results achievable using a slide window rather than a set of consecutive non-overlapping subsequences.

Acknowledgements

This work has been carried out in the framework of the ITINERIS2 project, Codice CUP F87G10000050009, under the financial support of Regione Lazio (Programme “Sviluppo della Innovazione Tecnologica nel Territorio Regionale”, Art. 182, comma 4, lettera c), L.R. n. 4, 28 Aprile 2006).

Table 1. Results

<table>
<thead>
<tr>
<th></th>
<th>KTH</th>
<th>UCF sport</th>
<th>Youtube</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun et al. [17]</td>
<td>100.0%</td>
<td>86.9%</td>
<td>-</td>
</tr>
<tr>
<td>Wang et al. [20]</td>
<td>94.2%</td>
<td>88.2%</td>
<td>84.2%</td>
</tr>
<tr>
<td>Whole video</td>
<td>95.8%</td>
<td>86.7%</td>
<td>75.3%</td>
</tr>
<tr>
<td>MSC</td>
<td>97.0%</td>
<td>88.0%</td>
<td>84.4%</td>
</tr>
</tbody>
</table>

References