Analysis of temporal alignment for Video Classification

Katy Blanc, Diane Lingrand, Antonio Paladini, Luca Coviello, Dane Mitrev, Emily Söhler, Leonardo Guzman and Frederic Precioso
I3S Laboratory - CNRS/UNS/UCA
kblanc@i3s.unice.fr, Diane.Lingrand@unice.fr and Frederic.PRECIOSO@unice.fr

***Abstract*— Thanks to their sucess on image recognition, deep neural networks achieve best classification accuracy on videos. However, traditional methods or shallow architectures remain competitive and combinations of different network types are the usual chosen approach. A reason for this less important impact of deep methods for video recognition is the motion representation.**

The time has a stronger redundancy, and an important elasticity compared to the spatial dimensions. The temporal redundancy is evident, but the elasticity within an action class is well less considered. Several instances of the action still widely differ by their style and speed of execution.

In this article, we analyze the temporal dimension by focusing on its singular dynamism, and we focus on the normalization of temporal elasticity on sequences to reduce speed variation within a class. We propose a framework to temporally align video instance in a classification task using the latest temporal warping method, Generalized Canonical Time Warping (GCTW). We evaluate our strategy on video datasets where the intra-class variations lie in temporal dimension rather than in spatial dimensions. Finally, we show the interest of accounting for temporal elasticity for a better video classification and we draw perspectives on more efficient ways to normalize simultaneously temporal and spatial intra-class variations.

I. Introduction

Current state-of-the-art methods for video classification are based on deep networks. The last decade has seen striking improvements in image classification thanks to the improvements of deep learning modeling, the increase of annotated visual contents and the fast development of computer power. In this context, the most intuitive strategy to extend these impressive results on images to video content has been to adapt the convolutional neural networks to the additional third dimension, the temporal dimension. Although deep architectures for images converged to state-of-the-art, there is not yet clear best video architecture topology for video classification. The current main deep networks for video can be described by three shared properties: (i) the convolutional kernels have two or three dimensions; (ii) the network input is the original RGB video, or the video optical flow or both (2-stream network); and (iii) the temporal information aggregation is made either by simple fusion (mean, maximum, etc.) or with a recurrent analysis. The best accuracies are generally obtained with deep neural networks with 2-stream approaches combined with 3D kernels and temporal aggregation [5], [23]. Consequently, there is no clear best architecture for all the video classification tasks, but each of them extracts diverse and meaningful information. One can see, by these different architectures and all the current works in the domain of video content representation, an effort to find a good way to characterize the time dimension. In CVPR 2017, a workshop entitled *"Brave new ideas for motion representations"* was opened in order to propose original ideas and to open discussions about motion, and the time dimension in video content.

978-1-7281-0089-0/19/$31.00

The time has two main properties: redundancy and elasticity. The temporal elasticity named the fact that an action is executed with a variation of speed and style. Several papers focus on reducing the time redundancy without taking into account the elasticity. Recently, Wang *et al.* proposed the Temporal Segment Network, a 2-stream network with a uniform sampling on the temporal dimension on each input video [23]. Thus, the temporal dimension is intermittent and the motion cannot be fully described. On the last ActivityNet Challenges, the less recognized classes represent motions as washing face and rock-paper-scissors, and the most recognized classes have a clear static visual cue as camel ride and ice fishing [6].

In this paper, we build a framework in order to classify video actions by focusing on the motion. We use an extension of the Dynamical Time Warping (DTW) to align actions on the same speed and thus reduce the intra-class temporal elasticity. Then we train and classify the aligned videos using a 3D deep network. In the next section, we present several applications of DTW or its adaptation in the video classification domain. In section III-A, we will present GCTW, the extension of DTW for a set of sequences. Finally, we will present our experiments to normalize the speed, and the improvement of the action classification thanks to this alignment.

II. Related works

Nowadays, Dynamic Time Warping is applied in diverse domains as computer graphics or bioinformatics [4], [1]. However, DTW has originally been applied on speech recognition [14]. DTW is an algorithm to temporally align a pair of sequences by optimizing temporal warps to maximize the similarity between them. The output of a DTW process is then a correlation score and the two temporal warping paths to obtain the aligned sequences. Some other methods like HMM, RNN or Action Spectogram [15], [3], [13] temporally compare sequences, however, their use is made more for sequence segmentation and matching based on vocabulary than for temporal alignment.

The computed correlation in the DTW algorithm depends on the representation's element and the chosen distance. As for the temporal warping, the representation and the distance can also be optimized by DTW in order to increase the correlation and thus, reduce the intra-class variation. Optimizing the representation to this end is equivalent to make this representation specific for common content and ignoring individual information. Hsu *et al.* [9] suggest keeping residual information to retain the pace of the individual style of a motion. They present the Iterative Motion Warping (IMW) method, which alternates between time warping and spatial transformation. Junejo *et al.* [11] use alignment to match two viewpoints of the same scene and find a robust descriptor to viewpoint changes. Spatial transforms and complex representations have then emerged, in particular, thanks to the introduction of spatiotemporal manifold model (STM) to align 3D motion capture data and the associated geodesic distance [20], [7]. Consequently, DTW is mainly used for searching robust representations.

After being adapted to specific cases, DTW was also extended to more generalized cases. First, following the idea of IMW to introduce an alternated spatial transform to the temporal transform, DTW has been naturally extended thanks to the Canonical Correlation Analysis (CCA), which linearly projects the representations into a common latent space: Canonical Time Warping (CTW) [26]. This linear projection manages the impact of each feature in the correlation estimation. More generally, CTW is multi-modal because each sequence has its own spatial projection, and thus, two different representations can be projected into a common space to compute their similarity. Then Trigeorgis *et al.* [19] propose a spatial transform made by a neural network to extend spatial projections to non-linear transforms.

While the alignment is still made by pairs in CTW and DCTW, Zhou *et al.* present the Generalized Canonical Time Warping (GCTW) [27]. This approach models the correlation on a set of sequences by the sum of the correlation of each pair, inspired by mCCA [8]. Moreover, GCTW uses a Gauss-Newton temporal warping, parametrized by a monotonic function basis.

Consequently, DTW and its extensions are useful methods to match common elements (the action, the scene) and reduce individual information (the view, the style, the speed). As a direct consequence, several tasks can take benefit from the reduction of these intra-class variations. Wang *et al.* [22] put more weights on early matching in GCTW (TCTW) in favour of a fast decision to predict the action with a k-NN classification. In [18], they combine DCTW with LDA, called DDATW, to align their temporal labels with each video in a temporal annotation task.

In this article, we present a generic video classification framework based on GCTW and we analyze the impact of temporal warping on the classification accuracy.

III. Classification with alignment

A. GCTW for temporal alignment

For the temporal alignment, we choose the Generalized Canonical Time Warping (GCTW) [27] for its multi-sequence alignment capacity. Given a collection of m time series, $\{X_i\}_{i=1}^m$, GCTW searches for all sequence X_i of length n_i, $X_i = [x_1^i, \cdots, x_{n_i}^i] \in \mathbb{R}^{d_i \times n_i}$, a linear spatial transform $V_i \in \mathbb{R}^{d_i \times d}$ and a non-linear temporal transform $W_i = W(p_i) \in \{0,1\}^{n_i \times l}$ parametrized by $p_i \in \{1 : n_i\}^l$, such that these output series $V_i^T X_i W_i \in \mathbb{R}^{d \times l}$ are aligned all together. GCTW minimizes the sum:

$$\min_{\{V_i\}_i \in \Phi, \{p_i\}_i \in \Psi} J_{gctw} = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{1}{2} \|V_i^T X_i W_i - V_j^T X_j W_j\|_F^2 + \sum_{i=1}^{m} (\phi(V_i) + \psi(p_i)) \tag{1}$$

where F indicates the Frobenius norm, $\phi(\cdot)$ and Φ are respectively the regularization term and space of the spatial transform V_i ans $\psi(\cdot)$ and Ψ are the regularization term and space of the temporal transform W_i (see [27] for details).

B. Video Alignment

We aim at reducing the speed variation within a class, and we choose GCTW for its capacities to simultaneously align a set of sequences, its spatial transform to select features and its parametrized temporal mapping.

As we have seen in the related work, it is necessary to choose a representation to apply DTW extensions. Due to the video redundancy and the computational cost, we use a representation framework inspired by the challenge YouTube 8M [2]. The frames are described using Google Net [16], and then the features are reduced using a PCA. Using the PCA features and the Euclidean distance, we align all video sequences within each class. In this way, we only use the deep representations to compute the temporal transform W_i by GCTW and we apply them directly on the video frame sequences (illustrated in figure 1). Thus, we end the process with the original video (of different lengths n_i) and the aligned video (of length l).

C. From alignment to classification

Using GCTW and our framework, we can assume that the classification problem can benefit from this class variation reduction. Then, given a database with C class, each class having m videos, we apply GCTW on each group of m videos within each class to get aligned videos. Thus, the GCTW framework gives us a database with no temporal elasticity within a class. We can now train our classifier on this new database. In the experiments, we will use the C3D Network as our baseline classifier, but any type of video classifier could be considered. In the experiments, we analyze the impact of the temporal alignment on this classification baseline through different testing protocols.

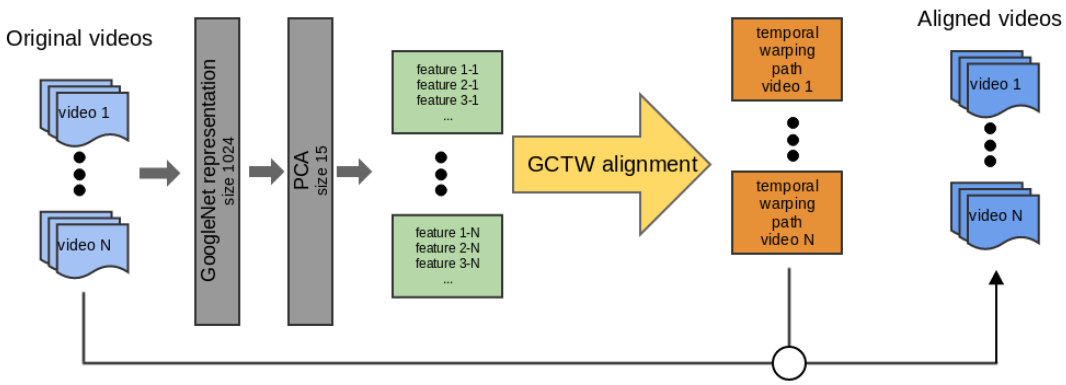

Fig. 1. Alignment Framework using GCTW within one class.

IV. Experiments

A. Data

To focus on speed normalization, we choose a database that contains low intra-class variation except the speed variation: the American Sign Language ASL [24]. The ASL database is composed of 1204 RGB videos of size 320×240, with $n_i = 20$ to 170 frames, illustrating 43 different signs. Although there are 14 different subjects, two illumination directions and some silent video parts, the intra-class spatial variations are actually low in this base (Fig. 2).

Conversely, the intra-class temporal variation is significant in this database. Figure 3 shows the video duration distribution per class. If we neglect the silent frames, the length is a good indicator of the speed variation inside a given class. This database is suitable to analyze the temporal elasticity as the major intra-class variation.

We also consider the IsoGD database [21], which is composed of 47933 videos of 21 subjects making 249 gestures. We only considered the RGB frames like for ASL. This database contains several languages: deaf language, diving language, Italian expression signs, etc. We can thus evaluate our framework with diverse class instances for the alignment and more sample per class for the training.

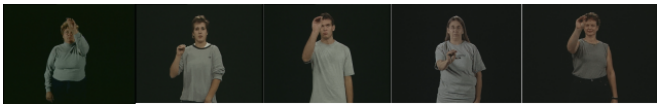

Fig. 2. Example of images from the ASL database: fives subjects doing the same sign.

B. Alignment and classification

Frames are represented by the 1024 length feature output from Google Net [16]. Then we reduce the dimensionality of feature space using a PCA with $d_i = 10$ coefficients (preserving 92% of the variance). For the classification step, we choose to use the C3D deep network [17] for its capacity to extract discriminative spatiotemporal patterns from learned three-dimensional filters. The classification network C3D is one of the admitted state of the art approaches for video classification.

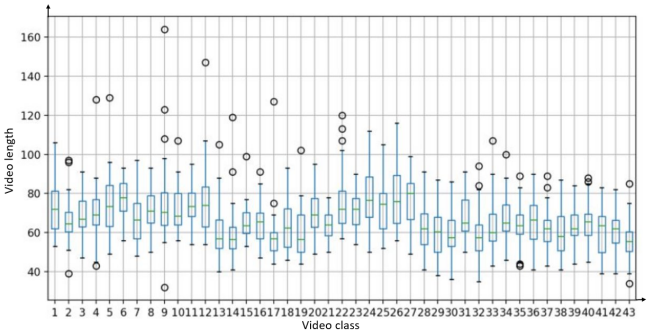

Fig. 3. Statistical box plot diagram showing the video length distribution within each class. The video length is usually around 50 and class 9 contains both the longest and the shortest videos.

Fig. 4. Images from the IsoGD database: four subjects doing the same sign.

The C3D Network takes as input fixed-length videos[25]. To normalize all videos' length while preserving the original speed distribution, either we pad each short video with the last frames until 140 frames or we crop the 140 centred frames of a long video. C3D does not learn filters from padding frames since static parts should not be discriminant. Thus, we do not add noise in the baseline classification. For computational reasons, we uniformly reduce the frame rate per second, taking 1 frame on 4. On the illustration Fig. 5, one can see the shortest video on the first line after the length normalization and thus the sign is preserved. From now on, the normalized database will refer to this size-normalized database with the original speed distribution. In opposition, the aligned database will refer to the videos aligned from GCTW, with $l = 35$ to obtain the same size videos, and thus the aligned database is normalized by size and by speed.

In order to evaluate our classification framework, we have designed 4 protocols.

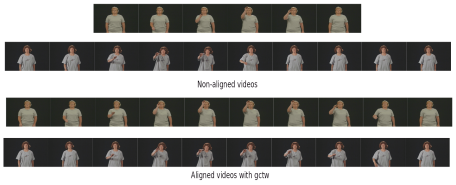

Fig. 5. Two video sequences of the same sign from the ASL database: before alignment at the top and after the alignment at the bottom. The first video is one of the shortest videos in ASL.

Classification protocol	ASL		IsoGD	
	top-1 acc	top-5 acc	top-1 acc	top-5 acc
Baseline	76.7	96.2	45	89.05
Protocol 1	14.5	37.9	41.65	71.03
Protocol 2	91.7	98.7	81.27	97.01
Protocol 3	50.4	92	81.34	97.11

TABLE I

TOP-1 AND TOP-5 ACCURACIES ON THE ASL AND THE ISOGD DATABASES ACCORDING TO TESTING PROTOCOLS.

*a) **Baseline**:* The baseline is our reference result for our study. C3D parameters are learned from the video database Sport 1-Million [12] and fine-tuned on the normalized database.

*b) **Protocol 1**:* C3D is trained on the aligned video samples (aligned samples) and test with diverse speed execution samples (normalized samples).

*c) **Protocol 2**:* C3D is trained and test on similar execution speed sample. This protocol is peculiar because we place the classifier in the ideal context to analyze the impact of this variation on the classification. Here we use the testing labels to align test data samples.

*d) **Protocol 3**:* Protocol 3 consists of training C3D on the aligned training set as for the two previous protocols. In testing time, each test sample is aligned to each class (already aligned samples). If there are C classes, then we get C temporal transforms W^c_{new} for one test sample. The prediction score of class c for a new sample X_{new} is computed as the prediction of class c for this new sample warped by an alignment with the class c. The predicted class is then the class with the maximum prediction score on the sample aligned with this class.

$$\mathcal{F}^c_{protocol3}(X_{new}) = \mathcal{F}^c_{C3D}(X_{new}W^c_{new}) \tag{2}$$

C. Results

The table I show the results for each protocol.

Firstly, we remark that protocol 1 falls in accuracy compared to the baseline on ASL and on IsoGD. Indeed, the network learns a sign always executed at the same speed but it sees new execution speed at testing time. A strong temporal variation between the train and the test set can thus hinder C3D recognition.

Secondly, with the protocol 2, we can remark that, in an ideal context without temporal elasticity within a class, the recognition rate gains 15% on ASL and 35% on IsoGd compared to the baseline. This implies that, if we build a robust representation for temporal elasticity, usual ConvNets can then be improved.

Finally, with protocol 3, we present a strategy to align samples without using the label in testing time. The results differ in ASL and IsoGD. The GCTW alignment is a generative method since it only considers one class during the process. Thus, it could either increase the similarity between different classes, as observed in the ASL database, or increase the discrimination between classes as observed in the IsoGD database. The alignment certainly reduces the discrimination in ASL because of the small inter-class variations and then, even if the sample is aligned to another class than its ground-truth, GCTW found enough correlations to warp it to this class. In the next section, we will discuss on the discriminative improvement of the temporal alignment.

V. DISCUSSION

DTW approaches have two main drawbacks. First, the generative aspect of DTW must be coupled with a discriminative criterion as we have seen in the last protocol. In DDATW [18], they combine DCTW and LDA to add a discriminative constraint over the spatiotemporal transform optimization. However, DDATW required temporal labelling which is a very specific context. Moreover, adding a discriminative criterion implies more sequence comparison and thus, increases the computational cost. Secondly, the alignment functions V^c_i and W^c_i depend on the sequence and on the class to be aligned with. This implies the computation of alignment during the test and with each existing class, to obtain all the aligned versions. One solution we are currently investigating is to learn the temporal alignment function T that can directly output the temporal warp W_i , independently from the class. Our preliminary work consists of extending the principles of the Spatial Transformer Network [10] to a temporal transformer network to learn a transform T directly from the classification labels. In this configuration, the temporal alignment T would not depend on the class and would optimize T in a discrimination aim.

VI. CONCLUSION

In this article, we have been focusing on the impact of temporally aligning videos in a video classification task. We have introduced a generic video classification framework combining GCTW for the alignment and C3D for the classification. Our system, and particularly the protocol 2, shows that a deep learning network can improve its accuracy by reducing the speed variation. When C3D is evaluated on an aligned database in train and test, the accuracy rate gain 15% from the baseline. We propose a particular protocol to make a prediction from the network trained on the aligned database, without knowing the class the sample has to be aligned with. Finally, we discuss the alignment and its drawbacks. In future works, we would adapt the Spatial Transformer Network to temporal alignment.

REFERENCES

[1] Aach, J., Church, G.M.: Aligning gene expression time series with time warping algorithms. Bioinformatics **17**(6), 495–508 (2001)

[2] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. CoRR **abs/1609.08675** (2016), http://arxiv.org/abs/1609.08675

[3] Böer, J.: Multiple alignment using hidden markov models. proteins **4**, 14

[4] Bruderlin, A., Williams, L.: Motion signal processing. In: conference on Computer graphics and interactive techniques. pp. 97–104. ACM (1995)

[5] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733. IEEE (2017)

[6] Chalenge, L.S.A.R.: Activity net challenge 2017. http://activity-net.org/challenges/2017/index.html

[7] Gong, D., Medioni, G.: Dynamic manifold warping for view invariant action recognition. In: ICCV. pp. 571–578. IEEE (2011)

[8] Hasan, M.A.: On multi-set canonical correlation analysis. In: IJCNN. pp. 1128–1133. IEEE (2009)

[9] Hsu, E., Pulli, K., Popović, J.: Style translation for human motion. In: ACM Transactions on Graphics (TOG). vol. 24, pp. 1082–1089. ACM (2005)

[10] Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: NIPS. pp. 2017–2025 (2015)

[11] Junejo, I.N., Dexter, E., Laptev, I., Perez, P.: View-independent action recognition from temporal self-similarities. IEEE trans. PAMI **33**(1), 172–185 (2011)

[12] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR , IEEE. pp. 1725–1732 (2014)

[13] McDuff, D., El Kaliouby, R., Kassam, K., Picard, R.: Affect valence inference from facial action unit spectrograms. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. pp. 17–24. IEEE (2010)

[14] Rabiner, L.R., Juang, B.H.: Fundamentals of speech recognition, vol. 14. PTR Prentice Hall Englewood Cliffs (1993)

[15] Sajjan, S.C., Vijaya, C.: Comparison of dtw and hmm for isolated word recognition. In: Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on. pp. 466–470. IEEE (2012)

[16] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al.: Going deeper with convolutions. In: CVPR. IEEE (2015)

[17] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: IEEE, ICCV (2015)

[18] Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S.: Deep canonical time warping for simultaneous alignment and representation learning of sequences. IEEE Trans. PAMI (5), 1128–1138 (2018)

[19] Trigeorgis, G., Nicolaou, M.A., Zafeiriou, S., Schuller, B.W.: Deep canonical time warping. In: IEEE, CVPR. pp. 5110–5118 (2016)

[20] Vu, H.T., Carey, C., Mahadevan, S.: Manifold warping: Manifold alignment over time. In: AAAI. vol. 1, p. 8 (2012)

[21] Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S., Li, S.Z.: Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In: IEEE , CVPR workshop. pp. 56–64 (2016)

[22] Wang, H., Yang, W., Yuan, C., Ling, H., Hu, W.: Human activity prediction using temporally-weighted generalized time warping. Neurocomputing **225**, 139–147 (2017)

[23] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE transactions on pattern analysis and machine intelligence (2018)

[24] Wilbur, R., Kak, A.C.: Purdue rvl-slll american sign language database (2006)

[25] Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. In: IEEE, ICCV. vol. 6, p. 8 (2017)

[26] Zhou, F., Torre, F.: Canonical time warping for alignment of human behavior. In: NIPS. pp. 2286–2294 (2009)

[27] Zhou, F., De la Torre, F.: Generalized canonical time warping. IEEE transactions on pattern analysis and machine intelligence **38**(2), 279–294 (2016)