Unsupervised Temporal Ensemble Alignment For Rapid Annotation

Ashton Fagg[1,2], Sridha Sridharan[2] and Simon Lucey[2,3]

[1] CSIRO, Brisbane, QLD, Australia
[2] Queensland University of Technology, Brisbane, QLD, Australia
[3] Carnegie Mellon University, Pittsburgh, PA, USA
ashton@fagg.id.au, s.sridharan@qut.edu.au, slucey@cs.cmu.edu

Abstract. This paper presents a novel framework for the unsupervised alignment of an ensemble of temporal sequences. This approach draws inspiration from the axiom that an ensemble of temporal signals stemming from the same source/class should have lower rank when "aligned" rather than "misaligned". Our approach shares similarities with recent state of the art methods for unsupervised images ensemble alignment (e.g. RASL) which breaks the problem into a set of image alignment problems (which have well known solutions i.e. the Lucas-Kanade algorithm). Similarly, we propose a strategy for decomposing the problem of temporal ensemble alignment into a similar set of independent sequence problems which we claim can be solved reliably through Dynamic Time Warping (DTW). We demonstrate the utility of our method using the Cohn-Kanade+ dataset, to align expression onset across multiple sequences, which allows us to automate the rapid discovery of event annotations.

1 Introduction

Time series alignment is an important problem for many areas of research - including speech processing, activity recognition, sensor networks and computer vision. Of particular interest is the alignment of time series which describe human motion. This problem is particularly challenging as the motions themselves may have disparate appearance. Such disparity may include differences in event speed and duration, physical differences between subjects and different presentation of the events themselves. These problems are amplified when considering the alignment of a set of sequences. If we were to attempt the alignment multiple sequences naively, a simple method would be to select a template from the sequences available and align all remaining sequences to that sequence. However this approach inherits several problems. For example, which sequence should be picked as a template? Can it be assured that this template produces reliable alignment across all of the sequences? This problem has been explored thoroughly in the image alignment domain, and has led to the proposal of methods known as ensemble alignment. For a set of semantically similar images, ensemble alignments aims to solve the alignment globally by finding a set of alignments which best align every image within the ensemble relative to all other images.

Taking the insights presented by image ensemble alignment [1] [2] [3] [4], and recent work in temporal alignment [5] [6], this paper will consider the application of ensemble alignment methodologies to multiple temporal sequences of the same modality. We propose that treating the set of sequences as an ensemble will enable an optimal alignment to be discovered, following the methodology proposed by [3] [4]. We make the assertion that semantically similar sequences, when aligned, should exist within a common, low rank subspace, which can be discovered using Robust PCA [7].

The alignment of a set of sequences lends itself to the automation of what is usually a tedious and time consuming task - event annotation. By solving for the alignment of the ensemble, we are able to rapidly discover event annotations for all sequences in the set. In this paper, we shall present an example which shows our method recovering approximate annotation for a set of sequences depicting facial expression onset.

1.1 Contributions

In this paper we shall present the following contributions:

- We present a novel framework for unsupervised alignment of an ensemble of temporal sequences.
- We demonstrate the use of RPCA [7] and DTW [8] to uncover a common low rank subspace for semantically similar temporal sequences.
- We demonstrate promising initial results on Cohn-Kanade+ for alignment of broad expression sequences for annotation of expression onset.

1.2 Notation Used In This Paper

Sets are notated as follows: $\mathbb{B}, \mathbb{R}$. Lower case bold letters denote column vectors. For example, an M dimensional column vector is denoted as $\mathbf{x}$, such that $\mathbf{x} \in \mathbb{R}^{M \times 1}$. Scalars are denoted by upper case non-bold letters. Upper case bold letters denote a matrix, e.g. $\mathbf{A} \in \mathbb{R}^{M \times M}$. Operations are denoted by a special font, e.g. the Lagrangian operator is denoted as $\mathcal{L}$, and the soft thresholding operator is denoted as $\mathcal{S}$.

2 Prior Art

This section will review current literature in the areas of ensemble alignment and time series alignment techniques.

2.1 Ensemble Alignment

Ensemble alignment, at its core, attempts to minimize misalignment over a set of samples. In the spatial domain, there has been significant interest in the area of multi-image alignment [1] [2] [4]. Of particular interest, is the RASL objective

[4] which decomposes the problem to be a set of simple problems, which are solvable using Augmented Lagrangian Methods. The motivation behind ensemble alignment is to exploit redundancies within the set of samples to recover a common, low rank subspace [7] in which all examples reside. The RASL objective [3] posits that an aligned ensemble of linearly correlated images can be formulated as:

$$\begin{aligned} \arg\min_{\mathbf{L},\mathbf{E},\mathbf{P}} \operatorname{rank}(\mathbf{L}) + \lambda\|\mathbf{E}\|_0 \\ \text{s.t. } \mathbf{D}(\mathbf{P}) = \mathbf{L} + \mathbf{E} \end{aligned} \tag{1}$$

where $\mathbf{L}$ describes a low rank subspace, $\mathbf{E}$ models sparse errors, and $\mathbf{D}(\mathbf{P})$ being the aligned ensemble, given a set of transformations represented by $\mathbf{P}$ and the original images $\mathbf{D}$. In effect, $\mathbf{L}$ describes a base image, where appearance variations are modelled by $\mathbf{E}$.

When solved using Augmented Lagrangian Methods [4], at each iteration the ensemble alignment problem is decomposed to set of discrete image alignment problems. Each image within the ensemble is warped with respect to the current estimate of the base image, solved using the Lucas-Kanade algorithm [9]. If one were to make the same assumptions about a set of similar time series, it is possible to posit the ensemble alignment problem in the same manner, where the temporal warping is discretely solved using proven time warping theory.

2.2 Time Series Alignment

In the area of time series alignment, there has been significant work based upon Dynamic Time Warping (DTW) [8]. DTW allows for the computation of a temporal warping which minimises the misalignment of two sequences. DTW is a powerful framework for time series alignment as it can be considered optimal when considering the distance between two sequences. The alignment path produced by DTW aims to reduce the distance between the sequences as much as possible.

If we define two 1D time series of different lengths, $\mathbf{x} \in \mathbb{R}^{N\times 1}$ and $\mathbf{y} \in \mathbb{R}^{M\times 1}$, the DTW objective which minimises the misalignment of $\mathbf{x}$ with respect to $\mathbf{y}$ can be formulated as:

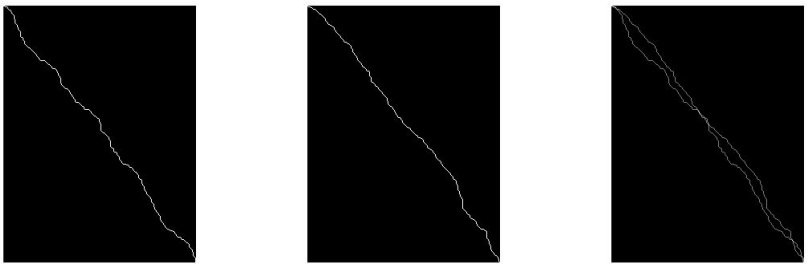
$$\operatorname{DTW}(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{P}\in\mathbb{B}} S(\mathbf{P}\mathbf{x}, \mathbf{y}) \tag{2}$$

Where, $\mathbf{P} \in \mathbb{B}$ encodes the alignment path between $\mathbf{x}$ and $\mathbf{y}$. The set $\mathbb{B}$ represents the set of all valid alignments, such that $\mathbb{B} \in \{0, 1\}^{M\times N}$. A valid alignment is defined as continuous and increasing in unitary increments. An optimal alignment can be efficiently drawn from $\mathbb{B}$ through the use of Dynamic Programming. S is a measure of cost, typically the least squares distance: $S(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$.

To understand the set $\mathbb{B}$, we visualise two examples of valid alignment paths computed using DTW. Figures 1a and 1b show alignment paths computed for random signals. In this instance, $\mathbf{P} \in \mathbb{B}^{320\times 240}$. However, one of the difficulties encountered in solving for valid alignment, is that the set $\mathbb{B}$ is non-convex. For a

convex set, it would be expected that any linear combination of valid elements of the set, would also lie within the set. For $\mathbb{B}$, this is assumption does not hold. In Figure 1c, the average of the paths shown in Figures 1a and 1b is illustrated. It is apparent the result does not lie with the set of valid alignments as the path is not causal and does not lie within the set of $\{0,1\}^{M\times N}$.

When combined with Augmented Lagrangian Methods, we assert that the use of DTW will ensure that optimal solutions can be drawn from $\mathbb{B}$ as DTW provides an efficient means of traversing the non-convex set and enforcing the alignment constraints.



(a) (b) (c)

Fig. 1: (a) and (b) illustrate random examples from the set of valid alignments, $\mathbb{B}$. Note that each of the paths is continuous, causal and increasing in unitary increments. In (c), we visualise the average of the two paths in (a) and (b) in order to demonstrate that $\mathbb{B}$ is non-convex. The path in (c) violates the constraints of $\mathbb{B}$ indicating non-convexity.

DTW features extensively in many existing time series alignment frameworks. Recent work in Time Series Kernels [10] [11] [12] highlights an important insight into sequences of common modality - that is, sequence similarity can be measured using DTW. This insight is applied in the formulation of the Time Series Kernel proposed by [11], which is in essence a measure of relative alignment. When applied to a temporal detection problem [12] it was shown that the Time Series Kernel provides excellent detection performance for broad expression detection on the Cohn-Kanade+ dataset [13]. Whilst the Time Series Kernel is in essence a measure of relative alignment between sequences, the representation is able to avoid a fundamental problem of temporal detection - variable event length.

Whilst DTW offers great utility to the temporal alignment problem, it has several drawbacks. Firstly, the alignment computed by DTW whilst considered optimal with respect to the pair of sequences, does not guarantee that the alignment will be meaningful. DTW makes the assumption that the sequence similarity is indicated by Euclidean distance. For many computer vision problems,

the Euclidean distance has been shown to be an unreliable measure of similarity due to the effects of corruption - for instance, a small amount of error in spatial registration of the features, or a cross-subject variability.

To compensate for these drawbacks, recent work has extended the DTW framework to incorporate tolerance towards sequence variability [14] [5] [6]. Canonical Time Warping (CTW) [14] attempts to account for cross-subject variability and mild amounts of sequence corruption by incorporating Canonical Correlation Analysis (CCA) into the objective. CTW parameterizes the temporal warping to a set of basis functions which maximise the correlation between the sequences. The work presented in [14] demonstrates superior performance when compared with regular DTW for sequences which are semantically similar, but contain variations in appearance.

Furthermore, CTW was extended to allow for the alignment of multiple sequences [5] of different modalities. The Generalized Time Warping (GTW) algorithm of [5] places emphasis on aligning sequences of different modalities. For example, given a sequence consisting of camera, motion capture and accelerometer data, the GTW algorithm is able to discern meaningful alignment of all modalities.

A further extension is the recent work proposed by [6] which intends to recover a common, low rank subspace for a pair of sequences which are corrupted by noise. For two sequences, a low rank projection is used to recover clean, aligned sequences from a pair of corrupted, but semantically similar sequences.

In this work we shall draw upon the insights from [4] [6], but apply them in a different manner. Rather than focusing on alignment of corrupted sequences and subsequent noise removal, we shall focus on using the power of DTW to uncover commonality on a larger scale - across many sequences where the definition of a reliable template may be difficult or impossible. In a similar manner to [4], we seek to minimize misalignment across a set of sequence by decomposing the global alignment problem to a set of independent alignment problems which are easily solved.

Using the alignment computed for the ensemble, we propose that the alignment can be used to rapidly generate sequence annotations, specifically for the onset of an expression. By aligning the whole set of sequences in time, we are required only to perform a minimal amount of annotation manually. Once the ensemble is aligned in time, in the best case all sequences will adhere to the same temporal profile, and expression onset can be annotated based on a single point in time.

3 Unsupervised Temporal Ensemble Alignment

Our method poses the ensemble alignment objective as an RPCA [7] problem:

$$\begin{aligned} \arg \min_{\mathbf{L},\mathbf{E},\mathbf{P}} \operatorname{rank}(\mathbf{L}) + \lambda \|\mathbf{E}\|_0 \\ \text{s.t. } \mathbf{L} + \mathbf{E} = \mathbf{D}(\mathbf{P}) \\ \mathbf{P}_i \in \mathbb{B} \ \forall i = 1, \ldots, N \end{aligned} \tag{3}$$

Where $\mathbf{L}$ describes a low rank subspace, $\mathbf{E}$ is the sparse error estimate and $\mathbf{D}(\mathbf{P})$ represents a set of DTW warps ($\mathbf{P}$) applied to the raw sequence ensemble ($\mathbf{D}$), such that:

$$\mathbf{D}(\mathbf{P}) = [\text{vec}(\mathbf{P}_1\mathbf{D}_1), \ldots, \text{vec}(\mathbf{P}_n\mathbf{D}_n)] \tag{4}$$

Each sequence, $\mathbf{D}_i \in \mathbb{R}^{F_i \times D}$ is warped to a predefined sequence length, F_0, by application of a temporal warping $\mathbf{P}_i \in \mathbb{B}^{F_0 \times F_i}$.

Similarly, $\mathbf{L}$ is defined such that:

$$\mathbf{L} = [\text{vec}(\mathbf{L}_1), \ldots, \text{vec}(\mathbf{L}_n)] \tag{5}$$

Where, $\mathbf{L}_i \in \mathbb{R}^{F_0 \times D}$. Hence, $\mathbf{D}(\mathbf{P}), \mathbf{L}, \mathbf{E} \in \mathbb{R}^{DF_0 \times N}$.

Equation 3 is considered difficult to solve efficiently due the non-convexity of the rank operation and L0 norm. Fortunately, a convex surrogate can be used in place of these operations to allow a solution to be found efficiently.

$$\begin{aligned} \arg \min_{\mathbf{L},\mathbf{E},\mathbf{P}} \|\mathbf{L}\|_* + \lambda\|\mathbf{E}\|_1 \\ \text{s.t. } \mathbf{L} + \mathbf{E} = \mathbf{D}(\mathbf{P}) \\ \mathbf{P}_i \in \mathbb{B}\ \forall i = 1, \ldots, N \end{aligned} \tag{6}$$

The substitution of the rank term for the nuclear (trace) norm enforces a convex lower bound on rank. We adopt the L1 norm to promote error sparsity.

This objective can be solved efficiently through the use of Augmented Lagrangian Methods (ALM). For purposes of simplicity, we express the ALM in scaled form [15]. The final objective is thus:

$$\begin{aligned} \arg \min_{\mathbf{L},\mathbf{E},\mathbf{P},\mathbf{X}} \|\mathbf{L}\|_* + \lambda\|\mathbf{E}\|_1 \\ \text{s.t. } \mathbf{L} + \mathbf{E} = \mathbf{X} \\ \mathbf{X} = \mathbf{D}(\mathbf{P}) \\ \mathbf{P}_i \in \mathbb{B}\ \forall i = 1, \ldots, N \end{aligned} \tag{7}$$

For purposes of simplicity, we expression the Lagrangian in scaled form [15]:

$$\mathcal{L}(\mathbf{L}, \mathbf{E}, \mathbf{X}, \mathbf{U}) = \|\mathbf{L}\|_* + \lambda\|\mathbf{E}\|_1 + \frac{\rho}{2}\|\mathbf{X} - \mathbf{L} - \mathbf{E} + \mathbf{U}\|_2^2 \tag{8}$$

where $\mathbf{U}$ are the scaled Lagrange multipliers, such that:

$$\mathbf{U} = \frac{1}{\rho} \times \mathbf{Y} \tag{9}$$

The algorithm can be summarized according to Algorithm 1.

3.1 Valid Solutions

For traversing the set $\mathbb{B}$, we assert that the use of DTW allows for an optimal solution to be gleaned for the alignment parameters. At each iteration, the problem of updating the ensemble alignment parameters is decomposed to individual alignment problems. Hence, we assert that utilizing DTW in a similar fashion to LK in [3] [4] allows for an acceptable solution to be found.

A large number of iterations typically allows a reasonable solution to be found. The following heuristics were used empirically to determine the feasibility of a solution:

$$\|\mathbf{X}^k - \mathbf{L}^k - \mathbf{E}^k\|_F < \alpha \tag{10}$$

where α is a small tolerance.

$$\|\mathbf{X}^k - \mathbf{X}^{k-1}\|_F == 0 \tag{11}$$

Data: $\mathbf{D}$ (raw ensemble), $\mathbf{P}$ (arbitrary time warps), $\mathbf{X} = \mathbf{D}(\mathbf{P})$
Result: $\mathbf{P}$ (optimal alignment paths), $\mathbf{X}$, $\mathbf{L}$, $\mathbf{E}$

Initialize $\mathbf{L}$, $\mathbf{E}$, $\mathbf{U}$ to zero matrices of appropriate dimensionality, $k = 0$, λ and ρ as appropriate.
while *not* ***converged*** **do**
  Update $\mathbf{L}$ using singular value soft thresholding:
  $(\Gamma, \Sigma, \theta) = \text{svd}(\mathbf{X}^k - \mathbf{E}^k + \mathbf{U}^k)$
  $\mathbf{L}^{k+1} = \Gamma \times \mathcal{S}_{\frac{2}{\rho}}[\Sigma] \times \theta^T$
  Update $\mathbf{E}$ using soft thresholding:
  $\mathbf{E}^{k+1} = \mathcal{S}_{\frac{2\lambda}{\rho}}[\mathbf{X}^k - \mathbf{L}^{k+1} + \mathbf{U}^k]$
  Update $\mathbf{P}$ and $\mathbf{X}$ using DTW:
  $\mathbf{P}_i^{k+1} = \text{DTW}(\mathbf{D}_i, \mathbf{L}_i^{k+1}))\forall i = 1, \ldots, N$
  $\mathbf{X}^{k+1} = \mathbf{D}(\mathbf{P}^{k+1})$
  Lagrangian update:
  $\mathbf{U}^{k+1} = \mathbf{U}^k + \rho(\mathbf{X}^{k+1} - \mathbf{L}^{k+1} - \mathbf{E}^{k+1})$
  $k = k + 1$
end

Algorithm 1: Algorithm for Unsupervised Temporal Ensemble Alignment for Rapid Annotation.

4 Experimental Evaluation

4.1 Performance Metrics

There are two key areas of performance which were considered for the evaluation of this work. First, we considered the number of sequences within the ensemble

which are aligned to ground truth at a given point in time. The point at which the most sequences correspond to their ground truth frame is considered to be the "consensus" point. To evaluate the performance over the set of sequences we define a measure of global alignment. We define x to be an alignment tolerance threshold, and evaluate the error for a given threshold as:

$$\mathcal{E}(x) = \frac{\text{Number of sequences which are within } x \text{ frames of ground truth}}{\text{Total Sequences}} \quad (12)$$

The overall performance measure is the area under the curve produced when x is varied from 0 (aligned) to the maximum possible misalignment. Misalignment is measured with respect to the target ensemble - that is, we evaluate the misalignment given the consensus point and the first appearance of the ground truth frame within the aligned sequence.

Using $\mathcal{E}$, we define qualitative measures of performance for each of the methods. These qualitative measures are to allow for a small degree of tolerance for misalignment across the set. These measures are:

- Perfect - Sequence matches ground truth exactly (highlighted below in green).
- Acceptable - No more than 10% of F_0 error (highlighted below in blue).
- Critical - More than 10% of F_0 error (highlighted below in orange).

An indicator of good performance would be a large number of "Perfect" alignments, with no "Critical" alignments. Bad performance would be indicated by the presence of "Critical" errors, no matter how many "Perfect" alignments are presented.

4.2 Cohn-Kanade+

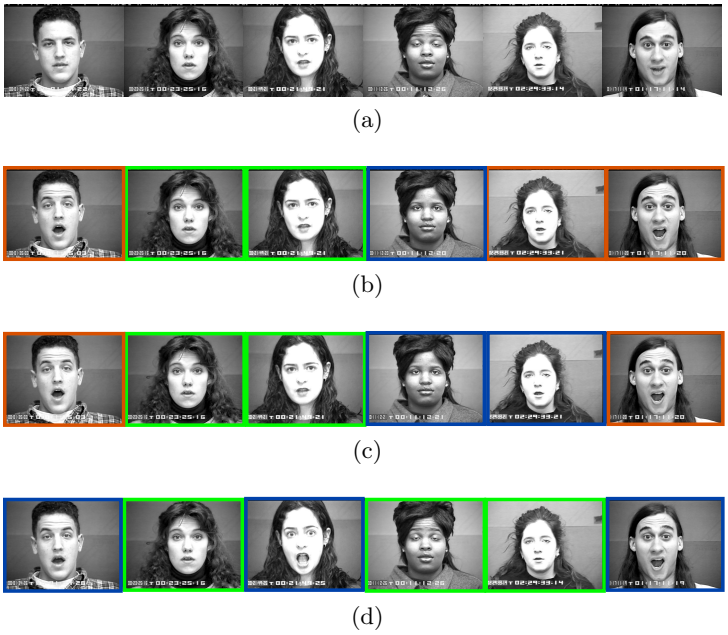
For evaluation, we used the Cohn-Kanade+ dataset [13] for aligning ensembles of sequences which are labelled as the same expression category. We utilised spatially normalised 2D landmark data which describe the appearance of the face. All 68 landmarks were used. As the sequences are of different lengths, we compute an initialisation for each sequence which consists of a random path computed using DTW to initialise the sequences to the chosen ensemble length (100 frames). The random initialization was used so as not to bias the initial alignment in favour of any particular sequence and to demonstrate worst-case performance where no sequence annotation is provided. For evaluation, the emotion categories of "Anger", "Surprise" and "Disgust" were selected.

The proposed ensemble method was evaluated against two sequence to template techniques, DTW and CTW. As a template, we randomly selected an initialised sequence from each class. Subsequently, we aligned all sequences in the class to this selected template using each method.

For the ensemble method, we initialised using the strategy above and aligned all sequences within each category.

For all three methods, we randomly selected a subset of sequences for evaluation and manually annotated the onset of the expression. We used this ground

truth to evaluate the "unsupervised" alignment. To ensure integrity of ground truth selection, the subject selection and annotation was undertaken separately to the evaluation of alignment results.



(a)

(b)

(c)

(d)

Fig. 2: Results from Cohn-Kanade+ Surprise category. Error categories are indicated for each sequence by green (Perfect), blue (Acceptable) and orange (Critical). (a) Ground truth. (b) Sequence to template alignment computed with DTW. (c) Sequence to template alignment computed with CTW. (c) Alignment computed using the ensemble method. Note that the ensemble method perfectly aligns 3 sequences (as opposed to 2) and produces acceptable alignment for the remaining sequences. Meanwhile, both DTW and CTW contain alignment errors which can be deemed critical.

The entire "Surprise" category consisting of 83 sequences was reduced to a rank 3 basis by our ensemble method. The ground truth for the six sequences evaluated is shown in Figure 2a. In Figures 2b and 2c, it can be observed that there is little correspondence across the sequences. Most sequences are behind the ground truth, with DTW and CTW presenting the most error - a maximum of 82 frames and 74 frames respectively. Our method (shown in Figure 2d) recovers optimal synchronisation for three of the sequences. Whilst alignment is not consistent across the selected samples, the maximum error present is 6 frames, within the defined tolerance. The error curve shown in Figure 3 shows that the ensemble method (shown in green) offers superior performance than

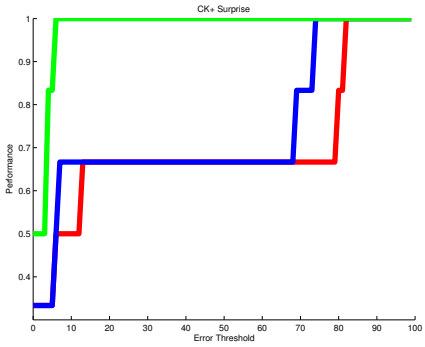


Fig. 3: Error curve for Cohn-Kanade+ Surprise category. DTW in red (AUC = 0.691), CTW in blue (AUC = 0.733), ensemble method in green (AUC = 0.969).

DTW (red) and CTW (blue), not only in accurately synchronising the most sequences but also in minimising the misalignment for the entire ensemble.

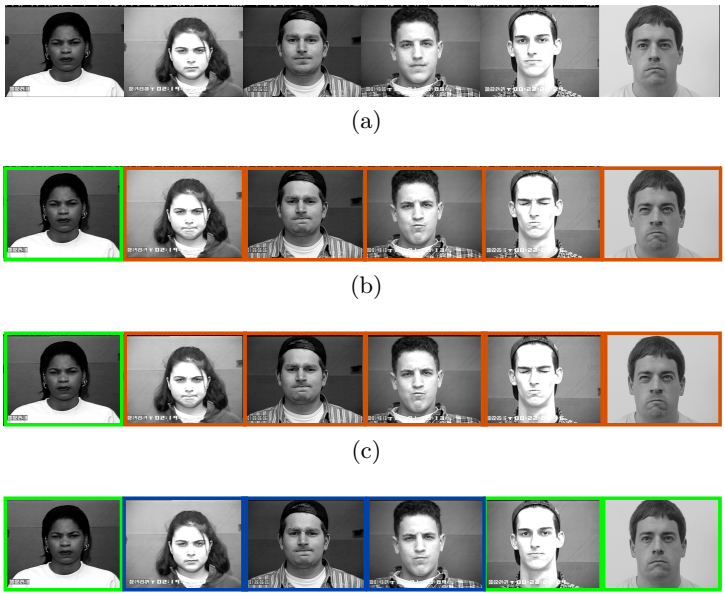
The "Anger" category consisting of 45 sequences was reduced to a 3 basis by our method. The ground truth for the six selected sequences evaluated is shown in Figure 4a. In Figures 4b and 4c, it can be observed that there is little consistency across the sequences in terms of expression progression with respect to ground truth. Both methods do not successfully align any of the ground truth frames, and return a maximum error of 86 (DTW) and 82 (CTW) frames. The results from the ensemble method are shown in Figure 4d, which show three ground truth frames in correspondence. Whilst the other sequences are not in correspondence, the maximum error returned by the ensemble method is 5 frames. The error curves shown in Figure 5 shows the ensemble method (shown in green) offers superior performance to both sequence to template methods.

The "Disgust" category, consisting of 59 sequences was reduced to a rank 4 basis by our method. The ground truth for the six selected sequences is shown in Figure 6a. In Figure 6b and 6c, both DTW and CTW have successfully aligned two of the sequences in accordance with ground truth. However, both sequence to template methods have significant error across the remaining sequences - 90 frames for DTW and 76 frames for CTW. The ensemble method also successfully aligned two of the sequences, with a maximum observed error of 4 frames.

5 Discussion

5.1 Alignment Consistency

The experiments performed highlight the effectiveness of our method over template based methods. The ensemble alignment outperforms sequence to template alignment for all three selected CK+ categories. Whilst the number of sequences in alignment after processing may not necessarily be greater than the alignment



(a)

(b)

(c)

(d)

Fig. 4: Results from Cohn-Kanade+ Anger category. Error categories are indicated for each sequence by green (Perfect), blue (Acceptable) and orange (Critical). (a) Ground truth. (b) Sequence to template alignment computed with DTW. (c) Sequence to template alignment computed with CTW. (c) Alignment computed using the ensemble method. The ensemble method successfully aligned three of the sequences with those remaining being in acceptable alignment. DTW and CTW do not reach a consensus point, and all sequences are critically misaligned.

recovered by template based methods, the overall misalignment across the sequences is greatly reduced (often by an order of magnitude). This results in a set of sequences which are vastly more synchronised.

Across all three emotion categories, the ensemble method performs adequately, with no errors in the "Critical" category. Raw CTW and DTW, whilst able to recover adequate alignment in some instances, encounter some "Critical" errors. In the case of Surprise and Anger, the ensemble method returns more "Perfect" alignments.

Disgust, however, yields interesting results across all three methods. For all three methods, a maximum of 2 "Perfect" alignments are returned. However, it is of note that the ensemble method does not encounter any "Critical" levels of error, whilst both DTW and CTW yield some "Critical" errors. Disgust is considered to be a more difficult category, as the presentation of the expression is more varied than other expression categories.

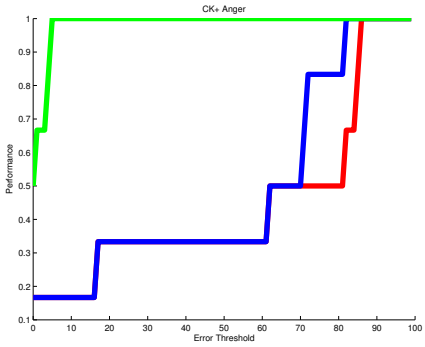


Fig. 5: Error curve for Cohn-Kanade+ Anger category. DTW in red (AUC = 0.441), CTW in blue (AUC = 0.488), ensemble method in green (AUC = 0.976).

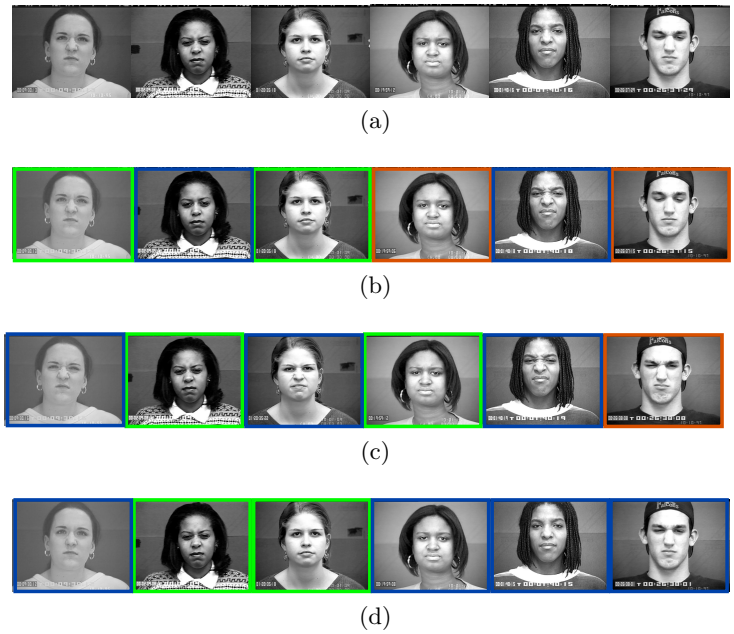
Whilst not perfect, the ensemble method is shown to outperform DTW and CTW for approximating expression onset annotation. It is possible that a different initialisation strategy for our method may result in better performance. For example, rather than initialising every sequence against a random alignment path, if a subset of the ensemble was correctly aligned a priori, this may be sufficient to boost alignment performance over the entire ensemble.

5.2 Convergence & Scalability

Convergence of large ensembles (such as those representing an entire category of CK+) typically occurs within 10,000-15,000 iterations (a few hours on a single CPU using MATLAB). However, convergence of smaller ensembles can occur within a few hundred iterations. It is worth noting that the scalability of the algorithm may be affected as the number and length of sequences grows. It is possible that modification of the objective as demonstrated in [16] may improve performance.

6 Conclusion

In this paper, we have proposed an ensemble-based approach for the alignment of semantically similar time series and its application to the discovery of approximate event annotation. Through the use of Dynamic Time Warping, we have demonstrated the application of insights from image ensemble alignment to be reasonably effective in the time domain. The proposed method delivers promising results for alignment and annotation generation of sequences consisting of facial expression onset.



(a) (b) (c) (d)

Fig. 6: Results from Cohn-Kanade+ Disgust category. Error categories for each sequence are indicated for each sequence by green (Perfect), blue (Acceptable) and orange (Critical).. (a) Ground truth. (b) Sequence to template alignment computed with DTW. (c) Sequence to template alignment computed with CTW. (c) Alignment computed using the ensemble method. Note that the ensemble method does not return any critical levels of error.

7 Acknowledgements

This research was supported by an Australian Research Council (ARC) Discovery Research Grant DP140100793.

References

1. Learned-Miller, E.G.: Data driven image models through continuous joint alignment. PAMI **28** (2006) 236–250
2. Cox, M., Sridharan, S., Lucey, S., Cohn, J.: Least squares congealing for unsupervised alignment of images. In: CVPR, IEEE (2008) 1–8
3. Peng, Y., Ganesh, A., Wright, J., Ma, Y.: RASL: Robust alignment via sparse and lowrank decomposition. In: CVPR, IEEE (2010)
4. Peng, Y., Ganesh, A., Wright, J., Xu, W., Ma, Y.: RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. PAMI **34** (2012) 2233–2246
5. Zhou, F., De la Torre, F.: Generalized time warping for multi-modal alignment of human motion. In: CVPR, IEEE (2012) pp. 1282–1289

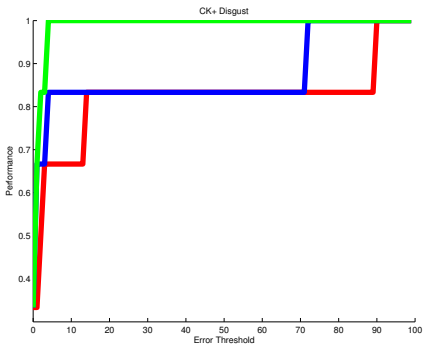


Fig. 7: Error curve for Cohn-Kanade+ Disgust category. DTW in red (AUC = 0.812), CTW in blue (AUC = 0.863), ensemble method in green (AUC = 0.98).

6. Panagakis, Y., Nicolaou, M.A., Zafeiriou, S., Pantic, M.: Robust Canonical Time Warping for the Alignment of Grossly Corrupted Sequences. In: CVPR, IEEE (2013) 540–547
7. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust Principal Component Analysis? JACM **58** (2011) 11
8. Müller, M.: Dynamic time warping. Information retrieval for music and motion (2007) 69–84
9. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision. In: IJCAI. Volume 81. (1981) 674–679
10. Cuturi, M., Vert, J.P., Birkenes, Ø., Matsui, T.: A kernel for time series based on global alignments. In: ICASSP. Volume 2., IEEE (2007) 413–416
11. Cuturi, M.: Fast global alignment kernels. In: ICML. (2011) 929–936
12. Lorincz, A., Jeni, L.A., Szabó, Z., Cohn, J.F., Kanade, T.: Emotional expression classification using time-series kernels. In: CVPRW, IEEE (2013) 889–895
13. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: CVPRW, IEEE (2010) 94–101
14. Zhou, F., De la Torre, F.: Canonical time warping for alignment of human behavior. In: NIPS. (2009) 2286–2294
15. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning **3** (2011) 1–122
16. Dai, Y., Li, H., He, M.: A simple prior-free method for non-rigid structure-from-motion factorization. In: CVPR, IEEE (2012) 2018–2025