Alignement temporel de gestes expressifs de communication

A. Heloir
Samsara Equipage Valoria
Université de Bretagne Sud
alexis.heloir@univ-ubs.fr

Concours Jeune Chercheur : Oui

Résumé

Among the different problems related to editing motion data, style translation has drawn a particular attention. In this paper we address the problem of motion temporal alignment applied to captured communicative gestures conveying different styles. We propose a representation space that may be considered as robust to the spatial variability induced by style. We then step on this motion representation to characterize the temporal variability between several realizations of a gesture sequence. By extending a multilevel dynamic time warping (DTW) algorithm, we show how this extension fulfill the goals of time correspondence between gesture sequences while preventing jerkiness introduced by classical time warping methods.

Mots clefs

Animation, conversational agents, gestural communication, style translation

1 Introduction

La conception et la mise au point d’un humanoïde virtuel capable de reproduire des gestes expressifs de communication avec la finesse et la dextérité qui caractérise le mouvement humain reste un problème ouvert dans la communauté de l’animation.

Le style selon lequel une séquence de gestes de communication est réalisée est porteur d’informations, d’une certaine manière enrichi le discours et aide à sa compréhension : le style est porteur d’indices d’ordres verbaux ou non-verbaux tels que les nuances dans le discours, l’intensité, les points d’emphase, le sexe du locuteur, son origine culturelle et son état émotionnel. Le style du geste étant partie intégrante de l’acte de communication, par conséquent, un système de génération automatique de séquences de gestes expressifs requiert que soit pris en compte la variabilité introduite pas le style au cours du mouvement.

Des travaux se basent sur des mouvements capturés pour proposer des cadres prenant le style du mouvement en considération. Parmi ces travaux, certains offrent à ‘’animateur des outils interactifs d’éditions et de manipulation du mouvement, ou des méthodes pouvant être intégrées dans des systèmes d’animation orientés données. Parmi ces approches, certaines décrivent le mouvement comme une combinaison de de composants représentant respectivement le style et le contenu intrinsèque d’une séquence gestuelle expressive. Notre démarche s’inscrit dans ce cadre, mais se focalise sur les gestes expressifs de communication. Nous scindons le geste en deux composantes : la première composante est supposée contenir la partie significative du discours, tandis que la seconde partie est supposée transmettre la partie d’ordre stylistique du mouvement.

Nous ne saurions poursuivre cet article sans donner une définition du style, définition sur laquelle nous allons nous appuyer durant e reste de cet article. Nous définissons donc le style comme étant l’a variabilité observée entre différentes réalisation d’une même séquence gestuelle (même contexte sémantique) exécutée selon des styles différents. Cette définition est volontairement de bas niveau car notre démarche est motivée par l’analyse du mouvement en tant que signal.

Les gestes de communication que nous avons capturé représentent différentes réalisation d’une séquence de langue de signe française (LSF). Pour capturer ces réalisations, nous avons demandé à un signeur professionnel d’interpréter une séquence déterminée de langue des signes en adoptant un style, une attitude, ou une vitesse d’exécution différente à chaque enregistrement.

La prise en compte de gestes expressifs de communication lève de nombreuses difficultés : une des principaux problèmes est du a au fait que ces séquences sont par nature non périodiques. Par conséquent, les séquences de gestes expressifs de communication vont sortir du cadre couramment traité par les précédentes approches d’édition de mouvement orienté, ces méthodes se basant sur des mouvements pseudo périodiques ou relativement courts tels que des séquences de locomotion ou d’art martiaux. Les gestes de communication présentent une caractéristique propres aux signaux expressifs : ils sont soumis au phénomène de coarticulation. Comme l’illustre la figure 1.

---

1 La coarticulation se manifeste par une influence importante entre les primitives du discours et rend leur identification ardue.
Ce travail se concentre sur l’identification de caractéristiques temporelles propres à différentes réalisations de gestes expressifs de communication. Pour ce faire, nous émettons l’hypothèse qu’il existe un mouvement fondamental commun aux différentes réalisations d’une séquence gestuelle donnée. Nous montrons ensuite qu’un espace de représentation s’appuyant sur une analyse en composante principales pondérée est compatible avec l’hypothèse d’existence de mouvement fondamental. L’alignement temporel est réalisé à partir de cet espace de représentation et optimisé par une approche multirésolution que nous appelons PCA adaptative. Nous montrons que cet algorithme d’alignement est capable de prendre en compte à la fois les variation temporelles d’influence locales et globales en le testant des réalisations d’une séquence de présentation d’un bulletin météo réalisé dans des styles différents.

The rest of this paper is organized as follow : in the first section, we present the existing works that address the problem of styled motion edition, then we introduce the style robust distance metric on which our Adaptive Fast Time Warping algorithm presented in section 3 steps on. Section 4 introduce the sign language CGS we acquired for the experiments detailed in section 5. Results are discussed in section 6. we then conclude by drawing perspectives of this work.

**2 État de l’art**

In order to introduce the problem of style retrieval and analysis from captured motions, we first show how model-based animation methods have somehow failed to produce yet natural and convincing motions. We then present works concerned by the characterization of style in motion sequences, and then give some insights on the temporal alignment problem between two motions.

**Méthodes d’animations basée modèles**

Le language signe a été étudié depuis les années 60 et a abouti à des systèmes de traduction/description dédiés [1]. Quelques systèmes de génération sont apparus oinspirés par ce paradigme. Lebourque & al. proposent un système de génération de gestes expressifs de communication où la tâche est exprimée comme une séquence de cibles discrètes dans l’espace euclidien autour d’un signeur virtuel. La séquence de posture est ensuite calculée par l’intermédiaire d’un système de cinématique inverse de type sensori-moteur. Plus récemment, le projet ESIGN [2] a mis en place un système de synthèse de gestes de la LSF guidé par un système de description et de la transcription de la langue des signes.

C’est à partir du XIXe siècle que des études du mouvement humain orientées styles sont apparues et ont été pour-suites jusqu’à aujourd’hui [3]. La théorie sous jacente a servi de base à des systèmes procéduraux de génération de mouvements humains [4]. D’autres systèmes procéduraux capables de prendre en compte le style sont apparus [5] en se basant des études psychocognitives [6]. Les méthodes d’animation procédurales se sont montrées capables de générer des animations intelligibles avec un haut niveau de contrôle, cependant, ces méthodes se sont révélées incapables de reproduire des mouvements humains dans toute leur naturalité et réalisme. Par conséquent, ont émergés d’autres travaux plus récents s’appuyant sur la capture de mouvements.

**Caractérisation du style basée données**


Amaya & al introduced the notion of *emotion* between a *neutral* motion and an *emotionally charged* motion [9]. By identifying the differences between two motions, the method is able to endow an arbitrary *neutral motion* with the emotion of the initial *emotional motion*. Several data driven motion generation methods exploit structured bases of motion segments [10, 11, 12, 13, 14]. By finding best sequence
of segments matching user defined tasks or constraints, the system is able to produce long motions with respect to some defined tasks/constraints. However, those methods do not directly address style editing, as the style of generated motion is bounded to the inner style of the captured motion. Stoneand & al. [15] combined psycho-cognitive knowledge and expressive motion data to set up an utterance generation system based on motion chunks concatenation. Motion generation is driven by a speech sequence. The resulting motion sequence then matches the prosodic aspects of the input speech sequence. Many of the style oriented motion edition/generation methods follow a common scheme which consists in extracting separately temporal and spatial features of styled motions. In the next subsection, we focus on temporal feature extraction issues.

Temporal characterization of style It has been well accepted that style affects both temporal and spatial characteristics of human motion. Thus, many motion editing methods address both the problem of temporal and spatial variance. As far as we know, those two concerns have always been addressed in a separate manner. Temporal and spatial aspects of style are characterized separately and recombined together during the motion generation step. Characterization of temporal stylistic features is by itself a non trivial problem and has been addressed, in most cases by relying on a well known non linear time alignment method: dynamic time warping (DTW). Bruderlin [8] adopted Sederberg’s shape based algorithm to vectors of postures described as Euler angles. Witkin & al. [16] adopted a motion curve based warping scheme over euler-angular posture vectors confining the motion editing process to a set of chosen articular trajectories. Rose [17] performed time warping over motion curves described as uniform cubic B-spline curves described by their control points. Gleichner [18] adopted dynamic time warping to align motions to perform relevant blending. Most recently, Shapiro [19] proposed an interactive motion editing tool that lets the user choose a relevant joint defined in Cartesian space to perform time alignment over two styled motions. Forbes & al. [20] introduce a motion search algorithm based on a weighted PCA-based pose representation. This algorithm evaluates a time warp distance between sequences of postures by performing bi-directional DTW from a seed point in a distance matrix built over the PCA-based representation space. The weighted PCA-based representation space introduced by forbes is of particular interest because it fits well with our needs of splitting a CGS into two parts: the fundamental motion that conveys the meaningful part of a gesture sequence and the stylistic content of the motion which conveys stylistic and emotional charge of the motion. This aspect is detailed in section 3. Although all the temporal alignment methods we described were proved to give satisfying results on motion sequence that are either cyclic (locomotion) or relatively short (martial arts moves) but give poor alignment results when applied to long realizations of a CGS. Handling this issue by splitting the CGS is not feasible as coarticulation blurs the frontiers between eventual motion primitives across different realizations of a CGS. The issue we encountered is actually well described by Keogh [21], who proposes an interesting discussion focused on the temporal aspects of human motion. He highlights the fact that temporal variability observed as both a local and a global influence, his work introduces the need of dealing with many different temporal variations. This leads us to design a new temporal alignment scheme that is multi-level in order to efficiently handle the multi-scale influence of temporal variations over several realizations of a long CGS. Our temporal alignment scheme is detailed in section 4.

3 Style robust representation space

It is difficult to formulate a precise definition of style as style is by many manners tightly mixed with captured motion data. An actor may perform a predefined motion sequence according to different moods, speeds, or expressivity clues, but, even when asked to be as neutral as possible, the actor will still convey its own kinematic signature. Still, if asked to perform a CGS, and provided that the actor can actually sign, each realization of the sign language gesture sequence will at least contain a common sub-part that will be sufficient to convey the meaningful part of the motion. The identification of this subpart motivates our investigations towards a low dimensional representation subspace for captured sign gesture data. The construction of a style robust distance function is motivated by the assumption that, for sign language, the meaningful part of the gesture is embedded in the subset which present the greatest variance. This assumption is in fact well illustrated when considering printed sign language lexicons. In such sign language manuals, lexicons are faithfully depicted by few images picturing the overall motion of upper the limbs and the handshapes.

3.1 Introducing Weighted PCA

The motion representation we designed to characterize the motion data in a reduced but still accurate orthogonal subspace is directly inspired from the work by Forbes & al. [20]. The weighted PCA-based representation they propose has the advantages of providing a coarse to fine representation that is driven by the amount of variance observed by each principal component in the original space. This representation is thus compatible with the assumption we made about motion properties. Furthermore, the weighted scheme introduced fits well with our requirement: the meaningful content of an expressive gesture sequence is mainly driven by the actors upper limbs, it thus make sense to introduce a weighting scheme that highlights arm and hands motion.

3.2 Motion data description

The motion data we deal with is composed by series of quite large vectors. As we deal with full body posture des-
cifications (hands + body), a posture vector is described by 63 unit quaternions. Our quaternion representation is then centered and linearized thank to the method presented in Johnson’s PH.D thesis [22]. This prepossessing step leads to a linear real valued representation of our posture sequences described by a matrix $M$ of rows 189 and columns, $n$, the number of frames in the CGS realization. Mean frame number of our CGS is 7000. Weights coefficients are then applied to each component of the posture vectors.

### 3.3 Eigenposture base extraction

PCA is a linear basis transformation that basically decomposes the original data so that any number of components accounts for as much as possible of the data variance. Mathematically, the principal components are the eigenvectors of the covariance matrix of the original data set. To perform PCA decomposition, we relied on the singular value decomposition (SVD) which, when applied to a reference PCA decomposition, we relied on the singular value of the covariance matrix of the original data set. To perform PCA decomposition, we relied on the singular value decomposition (SVD) which, when applied to a reference CGS realization matrix $M_{ref}$ leads to:

$$M_{ref} = U_{ref} \Sigma_{ref} V_{ref}^T.$$  

Where $V_{ref}$ and $U_{ref}$ are orthogonal unit matrices, $U_{ref}$ is an orthonormal eigenposture base of $\mathbb{R}^{189}$ whose $r$ first columns give the basis $u1, u2, ...$ of the optimal hyperplane of dimension $r$, $\Sigma_{ref}$ is a $189 \times n$ matrix with non-negative decreasing singular values on its diagonal. The, the $M_i$ subsequents realization of a CGS are projected onto the optimal basis extracted from $M_{ref}$.

$$V_i^T = \Sigma^+_i U_{ref}^T M_i$$

Where $\Sigma^+_i$ is the transpose of $\Sigma_{ref}$ with every nonzero entry replaced by its reciprocal.

This projection leads to a common representation space between every realization of a CGS. In this representation space, a realization $M_i$ of a CGS is described by the $r$ first rows of its matrix $V_i$. The distance between two poses is obtained by calculated the euclidian distance between their first $r$ scaled coordinates of $V_i$.

### 3.4 Projecting the motion onto an optimal subspace

Projecting an arbitrary motion on a eigenposture space is lossy if the projected motion is not included in the construction of the eigenposture basis. To minimize the error induced by the projection, forbes & al. constructed the motion search space on a eigenposture basis obtained from a motion sequence that conveyed the most variability: the range of motion (ROM) that had been used to calibrate the motion capture hardware. Such a choice is legitimate when no a-priori knowledge is available on the motion data to be projected.

Or we have a strong a-priori on the content of the motion data we deal with: each motion clip is a single realisation of a CGS. As a consequence, and considering the strong a-priori we have over the content of the data, it makes sense to build the projecting space upon a reference CGS that closely matches the motions we wish to compare. This decomposition leads to a more accurate representation space. In order words, the resulting representation space needs fewer eigenvectors to provide an accurate reconstruction of a pose belonging to the projected motion. Therefore, fewer coordinates are required to provide a faithful estimation of the fundamental motion. In our experiments, we found that taking $r = 4$, was sufficient to convince the overall meaning of any realization of a CGS. we thus define our distance function $\delta$ between two postures $q_i$ and $c_j$ belonging to two coordinates matrix $Q$ and $C$ expressed in the reference base $(U_{ref} \Sigma_{ref})$ as:

$$\delta(q_i, c_j) = \sqrt{(q_i(1) - c_j(1))^2 + ... + (q_i(n) - c_j(n))^2}$$

This distance function serves as a base to the construction of a distance matrix that will be handled by our multi-level adaptive DTW algorithm.

### 4 Dynamic Time Warping

Accurate matching of time series requires to take into account multiple levels of temporal misalignment, from uniform scaling along large sequences to small local misalignments. This fact has been identified as well in the biomechanical and the computer animation [21]. To handle this issue, we stepped on a coarse to fine approach which relies on a multilevel strategy that iteratively adjusts both the search space and slope constraints of DTW. By doing so, it becomes possible to avoid the trade-off between global and local adjustments. We show that this methods prevents discontinuous jumps from occurring during the Warp path extraction while preserving sufficient accuracy in time correspondence.

#### 4.1 Fast DTW

FastDTW algorithm has been introduced by Salvador & al. [23] and was initially designed to cut off the computational cost of the well known DTW [24], which is of $n^2$ in its standard implementation. FastDTW basically consists into splitting the complexity of standard DTW by recursively downsampling the time series to be warped. The warp path found at each iteration of the algorithm is then projected onto the higher resolution layer and serves a guide that reduces computational complexity by spatially reducing the area handled by dynamic programming, as illustrated in figure 3.
FastDTW complexity is $O(n^2)$, and is known to find an accurate minimum-distance warp path between two time series that is nearly optimal. Unfortunately, warp path obtained via FastDTW contain many consecutive horizontal or vertical step. This leads to introduce jerkiness, high discontinuities or long steady postures over warped gesture sequences.

### 4.2 Constrained DTW

Let $Q$ and $C$ be two time series of respective length $m$ and $n$. Let $\delta$ be a distance function between any element $q_i$ of $Q$ and any element $c_j$ of $C$. Let $k > 1$ be a real valued coefficient. Constrained DTW can then be recursively defined as:

$$
\gamma(q_i, c_j) = \delta(q_i, c_j) + 
\min(\gamma(q_{i-1}, c_{j-1}), k\gamma(q_{i-1}, c_j), k\gamma(q_i, c_{j-1}))
$$

$$
DTW(Q, C) = \gamma(q_m, c_n)
$$

The real valued coefficient $k$ prevents the warp path from leaving the main diagonal of the distance matrix between $Q$ and $C$. Constrained DTW limits the number of consecutive horizontal or vertical steps and provides a smoother matching. The drawback of constrained DTW is the slope limitation which is introduced. Slope limitation may prevent constrained DTW from finding a warp path that is faithful to the optimal warp path if the temporal variation has a too wide influence.

### 4.3 Adaptive DTW

Adaptive time warping was motivated by the wish of providing a constrained version of the DTW algorithm which could adapt to the many temporal misalignment levels observed in the realizations of CGS. To avoid the trade-off between warp path smoothness and a coarse to fine approach which relies on a multilevel strategy that iteratively adjusts both the search space and slope constraints of DTW. By doing so, it becomes possible to avoid the trade-off between global and local adjustments. We extended FastDTW algorithm in order to take into account an adaptive slope support at each iteration. The pseudo code in figure 2 describes the algorithm. The implementation is recursive and the base case occurs when one of the input motion becomes shorter than the window length we set in the parameters of Adaptive DTW.

## 5 Application

### 5.1 Communicative Gesture Sequence description

The earliest part of our work was dedicated to collecting sign language motion data performed according to various styles. To do so, we relied on an original experimental protocol which involved an optical motion capture system and a pair of CyberGloves. This choice was motivated by the fact that sign language involves frequent hand or fingers contacts, leading to visual occlusions which would be error prone when dealing with a only optical acquisition device.

We then asked a professional signer to perform several realizations of the same weather forecast presentation in French sign language (FSL) by adopting different styles. The mean duration of a sequence was 70 seconds and the CGS realizations were done according to the following style: neutral (two tries), emphasis, angry and tired. We took as reference realization the first sequence performed according to neutral style.

## 6 Results

We Preprocessed and merged the data to obtain one quaternions posture matrix per CGS realization. We then projected the motions on the eigenposture basis extracted from the reference CGS realisation. Finally, we applied standard DTW and our Adaptive DTW algorithm to the obtained matrices. Figures 4 and 5 depict the most evocative results. Figure 4 compares a temporal alignment obtained thanks to classical DTW and the temporal alignment obtained thanks to our Adaptive DTW. The plots show the evolution of the influence of the first eigenposture vector for each motion in the eigenposture representation space. The two motions are respectively the first neutral sequence and the second neutral sequence. Those sequences provide actually an interesting benchmark opportunity, because, on one hand, they are very close, since they were performed according to the same style, but, on the other hand, they suffer from heavy artifacts since the signer messed up a couple of signs between frame 2200 and frame 3800.

The second plot from top highlights the temporal alignment found by classical DTW. It appears that classical
DTW actually finds an optimal time warp path which minimizes DTW cumulative distance. Unfortunately, the obtained warp path introduce many discontinuities, especially in the messy part if the sequences. As a consequence, the warped motion obtained thanks to DTW contains many discontinuities. On the contrary, adaptive DTW shows its ability to find a smooth but still accurate warp path between the two motions. It as well reveals to be quite robust to artifacts.

Figure 5 depicts the time alignment obtained by aligning the angry styled CGS performance onto the reference CGS performance thanks to Adaptive DTW. The postures are equally sampled and the curves represent the evolution of the influence of the first the eigenposture vectors. This figure highlights the capability of Adaptive DTW to provide a smooth and accurate matching despite the spacial variations introduced by style.

7 Conclusion

Using communication gestures obtained through motion capture rise a number of difficulties that are inherent to the nature of these motions : difficult automatic or manual segmentation and strong variability between the execution of two realizations of a communicative gesture sequence (CGS). We have proposed in this article a motion alignment method that has proved to be robust to the temporal and spatial variabilities that are induced by differently styled realizations of a relatively long CGS. This information can be exploited in a variety of ways : motion editing, blending, segmentation or style translation. Our hypothesis lies on a possible decomposition of communicative gesture between a fundamental motion (common to all realizations) and a style content. This decomposition is obtained through weighted PCA decomposition methods. This decomposition is then used as input of an adaptive dynamic time warping algorithm which provide smooth alignment between sequences. Finally our method does not address the spatial registration between those gestures, and this point constitutes a challenging extension for our framework.

Références


