Extracting Higher-Level Information from Facial Mocap

J.P. Lewis*
Weta Digital

Ken Anjyo*
OLM Digital

Figure 1: a), b) automatic weight regions for skinning; c), d) importance of facial regions for emotion (c) and speech (d) discrimination.

Motion capture data is valuable information, but in a low-level form that makes subsequent editing operations difficult. In cases where extensive manipulation is needed a manually animated performance may be preferable, because the manually constructed representation has “artist-friendly” higher level controls. Algorithmic approaches can potentially extract higher-level and semantic representations from motion capture, but much work remains to be done in this area. In this talk we show examples of several types of higher level information that can be extracted using an affinity measure on the mocap data.

1 Affinity Measures

Ideally, we would like to discover temporally correlated movement regardless of the direction of that movement. Standard linear correlation is arguably a poor measure in this respect. For example, in a walking motion, the movement of one leg is more related to the movement of the other leg than to head motion, but linear correlation would not discover this because the leg movement is negatively correlated. Similarly, in making an “oo” sound in speech, the motion of the upper and lower lips is anti-correlated, while the inward motion of the lip corners is perpendicular to the up-down motion and thus has zero correlation.

We employ two measures. First considering only linear approaches, we seek a direction-invariant affinity measure. This can be formulated as maximizing the expected movement correlation of two mean-removed points $a, b$ by finding the rotation $R$ that best aligns their motion,

$$\arg \max_R \frac{1}{N} \sum a_i^T R b_i + \text{tr} \left( R R^T R - I \right)$$

where $L$ is a Lagrange multiplier matrix enforcing orthogonality of $R$. Taking the derivative and setting to zero, we have

$$RL = -\frac{1}{N} \sum a_i b_i^T$$

where the right hand side is given (note that it is an outer product, i.e. a “cross-covariance” matrix). The singular value decomposition $R L = U D V^T = (U V^T)(V D V^T)$ gives the solution, where $R = U V^T$ is the desired rotation taking $b_i$ into best alignment with $a_i$.

The direction-invariant linear measure handles the perpendicularly and anti-correlated examples mentioned previously. While we do not expect the motion of various points to have a perfect linear relationship, often the linear component of the motion is dominant. In these cases linear model is preferred due to learning and regression theory considerations of model simplicity and lower variance [1].

In other cases, however, a linear model is unlikely to be sufficient. To cover these cases we switch to an affinity measure based on the mutual information $I(X, Y) = H(X) + H(Y) - H(X, Y)$ between two variables $X, Y$, where $H()$ is the marginal or joint entropy. This measure reflects any functional dependence between quantities, i.e. the mutual information between $X$ and $Y = f(X)$ is high regardless of the nature of $f()$. Intuitively, mutual information is the information that a variable $Y$ provides about $X$ (or vice versa), measured as the expected number of bits that can be saved in specifying $X$ when $Y$ is known. Although motion capture data is not random, the mutual information concept can nevertheless be applied simply by treating normalized histograms of the motion capture variables as if they are probability densities, and then directly applying the mutual information definition.

2 Applications

Fig. 1 a), b) show the linear affinity measure applied to the task of determining weight map regions for cluster based skinning as implemented in systems such as Maya. The motion capture is obtained from a model-based video tracking system. The set of vertices that are significantly correlated with an indicated vertex (red) are highlighted in blue. Note that in a), the direction-invariant measure has discovered symmetric weight maps on both sides of the mouth, although the horizontal movement on one side of the mouth is typically anti-correlated with that of the other side. In b) the eyebrow movement was not completely symmetric.

Fig. 1 c), d) show the versatility the nonlinear affinity measure. Here $I(X, Y)$ is applied between point motion ($X$) on a face across images and subjective ratings ($Y$) of differences in emotional affect (c) and speech discrimination (d) across these images. For the emotional affect task, the cheeks are unimportant, the mouth region is slightly important, but the brow and inner eye regions are dominant – suggesting that motion capture markers should be placed in these areas, and further that a CG face model should carefully reproduce movement in these areas. In the speech task, the mouth is quite important, as expected, but the inner eyes are again more significant than one might expect. This result may be consistent with psychological research showing that people look at the eyes more than the mouth when listening to speech [2].

This work was partially supported by the Japan Science and Technology Agency, CREST project.

References
