Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than things you do expect) Plautus (ca 200(B.C.)

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE NO: D3.2
Setup and Stimuli for Neurophysiological Experiments

Date of deliverable:31.12.2006
Actual submission date: *29.01.2007*

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **K.U.Leuven**

Revision [1]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

DIRAC
Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

D3.2 SETUP AND STIMULI FOR NEUROPHYSIOLOGICAL EXPERIMENTS

Katholieke Universiteit Leuven (KUL)
Eidgenoessische Technische Hochschule Zuerich (ETHZ)

Abstract:

This deliverable describes the experimental setup and stimuli used in neurophysiological experiments for action recognition at KUL and a new set of stimuli that have been developed in collaboration between ETHZ and KUL.

Table of Content

1. Setup and Stimuli for Neurophysiological Experiments.

The setup to train the animals and record single cells from the animals is ready, as well as the custom-made software to show the images during a categorization task. The setup consists of a CRT display, custom-designed software to control the stimulus display, sample, display on-line and save eye position, control the behavioral task and fluid reward delivery. The animals are seated during the experiment in custom-designed monkey chairs and their head is fixed using a surgically implanted custom-designed head post. The electrophysiological signals are measured by microelectrodes. Their signal, after preamplification, is amplified and filtered and spikes are isolated with a custom designed spike discriminator. Spike times are saved by a DSP-based system. We have purchased a high temporal resolution (1000Hz) and spatial resolution system to measure the eye position of the animals during training and recording and this eye tracker has been successfully installed. Two rhesus monkeys have been implanted with a head fixing device and we have made extensive single cell recordings in these monkeys in the new set up.

For these recordings we have used a stimulus set developed in collaboration with Dr. F. Pollick (Univ. Glasgow, UK). This stimulus set consisted of three prototypical actions (throwing of an object, lifting of an object and knocking on a door; see **Figure 1** for 6 snapshots of the movie of the lift action) and their blends. The blending algorithm provided actions that consisted of mixtures of the three prototypes using different weights (20% steps) of each prototype (e.g. 0% throw, 80% lift, 20% knock; 20% throw, 60% lift, 20% knock, etc). This parameterized set of actions can be represented as a triangle with the prototypical actions at its corners and the different blends in between. The blending operation results in smooth transitions between the different action stimuli and allows a measurement of the tuning to the action stimuli in this action space (see **Figure 2** for an example of responses of a single cortical neuron to this set of movies). The action images were rendered as stick figures (Figure 1) and point light displays. In the latter type of stimuli, only point lights positioned on the limbs of an otherwise invisible agent (e.g. human) are presented, which, despite this highly reduced spatial and temporal information, can provide a strong perception of an acting agent. The objects of the goal-directed actions were not rendered. The responses of the neurons to these dynamic action stimuli were compared to responses to static images of individual snapshots, evenly sampling the action sequence. In addition, we could reduce the stimuli by systematically removing the lower limbs, trunk, etc of the actor to determine which part of the body configuration is necessary to drive the neuron.

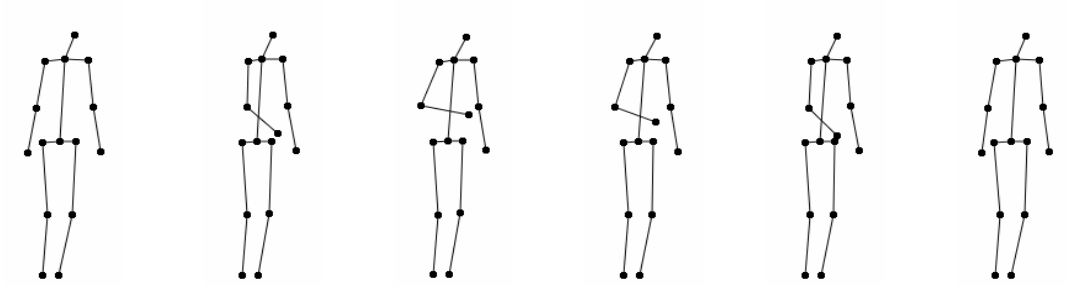


Figure 1. Selected snapshots of the movie depicting a person picking up a (not-rendered) object and putting it back.

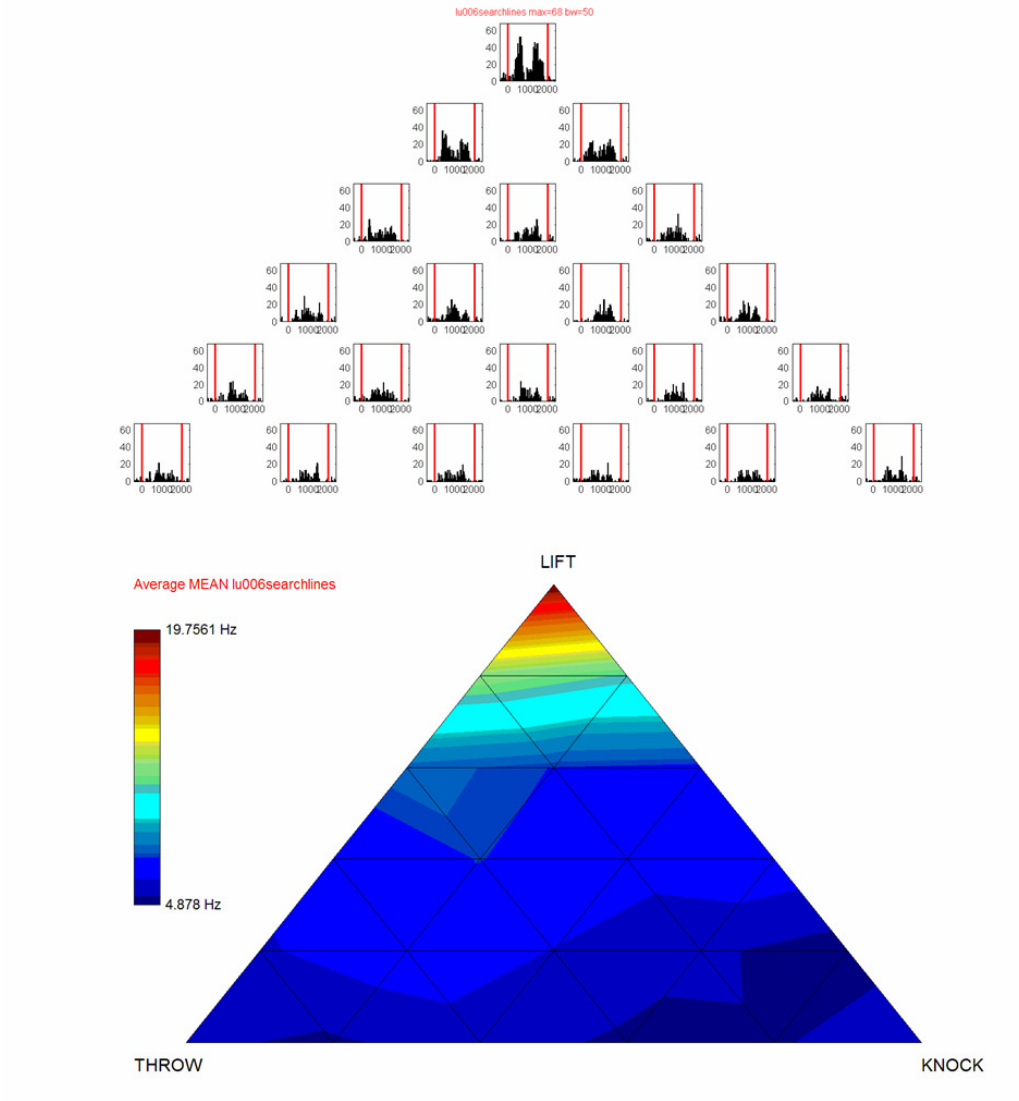


Figure 2. Responses of a single temporal cortical neuron to 21 actions of a parameterized action space. These actions consisted of the 3 prototypes (Knock, Throw and Lift actions) and their blends. The responses to the actions are represented in a triangular configuration with the corners corresponding to the prototypes. The stimulus duration was 2000 msecs. The responses are shown as PostStimulusTimeHistograms (upper panel) and as the mean firing rate (colour coded), averaged across the entire stimulus duration (lower panel). This neuron responded selectively to the Lift action and blended actions containing a strong Lift component. Note the bimodal response profile in the PSTHs indicating that this neuron responded at particular moments of the action sequence.

2. Second Stimulus Set

For the further neurophysiological and computational work in action recognition, a new database of motion sequences was created and processed. Both the neurophysiological (KUL) and computer vision (ETHZ) research tracks will focus on this corpus of data for their further investigation of the topic.

The dataset concentrates on the typical visual motion pattern that occur when humans move, more specifically on the most common types of human locomotion, *walking* and *running*. Multiple subjects were recorded under laboratory conditions performing the aforementioned activities at different speeds. The resulting three-dimensional motion data was then further processed, and transferred into representations that are suitable for the planned experiments in the neurophysiological and computational domain. On the one hand, we created visual stimuli that will eventually be presented to test subjects (macaque monkeys) in a single-cell recording setup. On the other hand, for the computational and statistical analysis of the data, we prefer a representation in terms of 3D joint trajectories or joint angles between rigid body limbs that will be used for training a statistical model for action recognition algorithms.

2.1. Motion Capture Setup

The sequences were recorded at the Motion Capture (MoCap) laboratory at ETHZ. This lab is equipped with an optical MoCap system (VICON) with 6 cameras that operate in the near-infrared range. In order to reconstruct the 3D body motions, 41 infrared-reflective markers are attached to the skin of the test subjects according to a specific protocol. The trajectories of these markers are then tracked in the individual camera streams and integrated into a 3-dimensional representation. Finally, an abstract body model (bi-ped, 17 rigid limbs) is used to interpret that data and solve for body poses. The system operates at 120Hz, and its spatial accuracy is better than 1cm. Such MoCap systems are often used in the movie industry, but also for medical purposes and biomedical analysis or therapy, and thus meet the demands of highly delicate applications. **Figure 3** illustrates the MoCap process.

The working volume of the setup is limited to approximately 2m by 2m, therefore a treadmill was used to allow for locomotion patterns of a certain duration (i.e. multiple running or walking cycles) without leaving the working area.

2.2. Human Action Sequences

Six subjects, male, between 20 and 40 years of age, of average physical constitution and in good health, were asked to perform a set of activities on the treadmill. They were allowed to acclimate to moving on the treadmill, which may be a bit cumbersome at first and which may else lead to biased or unnatural motions. The subjects were asked to walk and run at six different speeds for about 10 seconds. The speeds ranged from slow walking (2.5 km/h) over average speed to fast walking (4.2 and 6 km/h). Running was performed at 8, 10 and 12 km/h. The result of this stage are motion capture sequences (36, 6 subjects at 6 speeds), represented either as marker trajectories, or as a kinematic tree with 6 degree-of-freedom (DOF) transformations indicating the relative pose of each limb with respect to its parent limb, or with respect to a global coordinate system. The data is available in the following formats:

a) `*.c3d` raw marker data,
b) `*.V` kinematicly fitted body model (binary),
c) `*.csv` kinematicly fitted body model (text file).

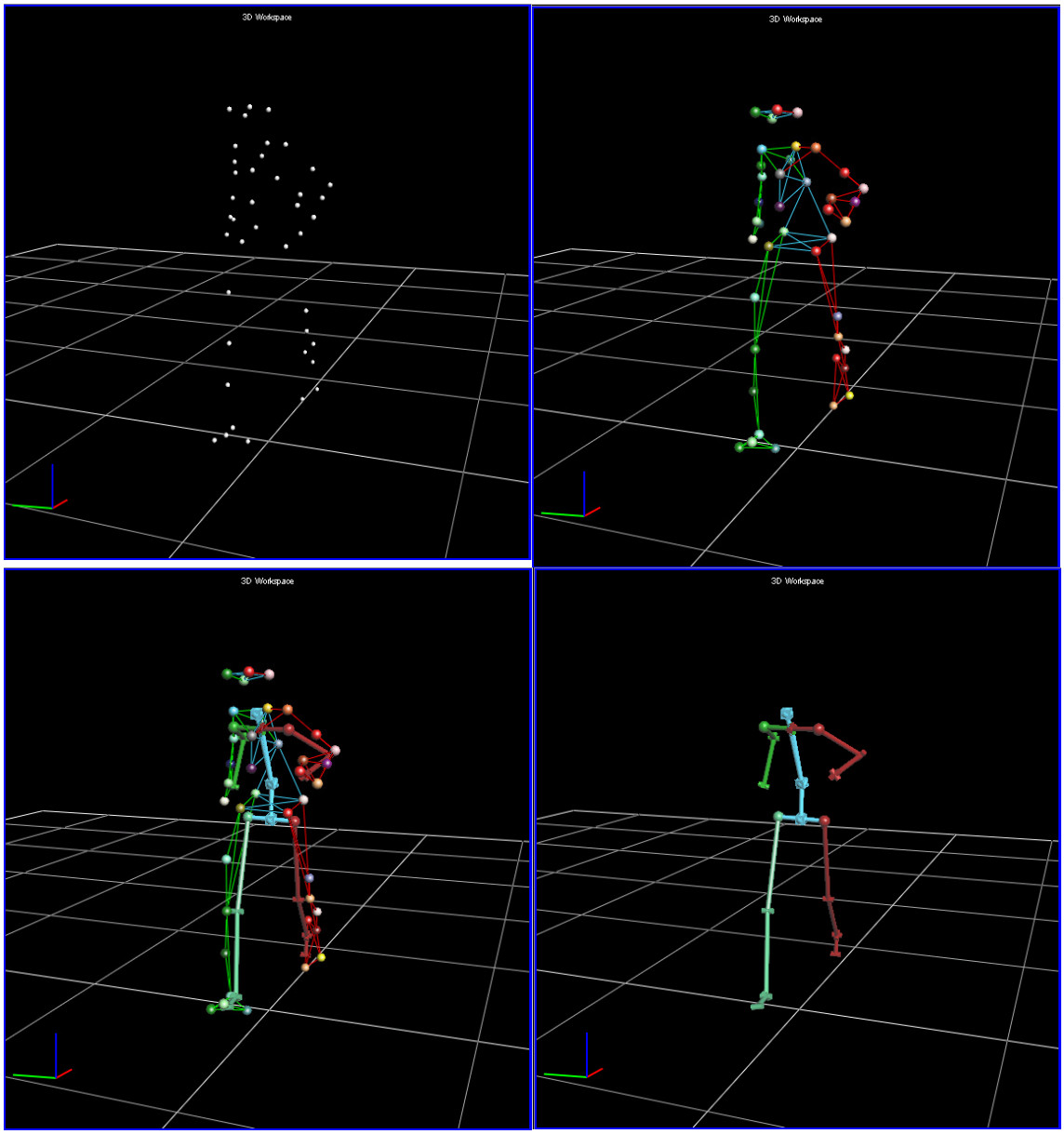
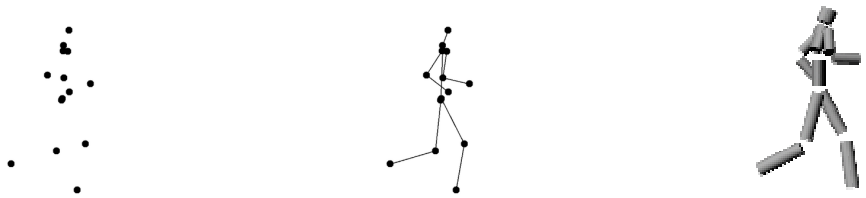


Figure 3. Visualization of the motion capture process: (top left) captured marker positions in 3D; (top right) connections between marker positions on the skin, from which the skeleton (bottom left) is inferred; (bottom right) final inferred skeleton.

2.3. Stimuli

For the neurophysiological work, a set of visual stimuli sequences was created. The idea here is to create reduced visual stimuli that allow for a better controlled analysis of the kind of pattern that are picked up the observer's brain than would be the case with real video recordings. Also, the abstract visual stimuli can be altered in a controlled and parametrized manner. Three different types of stimuli were chosen for this work:

a) point-light displays, were the joint positions are indicated by dots,
b) stick figures, where the adjacent joint are additionally connected by a line, and
c) c) humanoid figures, where body limbs are represented by cylinder-like geometrical primitives.



(a) (b) (c)

Figure 4. The three different types of stimuli prepared for neurophysiological experiments: (a) point-light displays where the joint positions are indicated by dots; (b) stick figures, where adjacent joints are connected by a skeleton line; (c) humanoid figures, where body limbs are represented by cylinder-like geometrical primitives.

Figure 4 shows example still images taken from the dataset. Each motion sequence was used to create one left-to-right and one right-to-left image sequence that will be used in a direction recognition or a backward/forward distinction task. The image sequences were processed at a frame rate of 60 Hz and a resolution of 640x480 pixels. For the point-light and stick sequences, bilinear interpolation was used to reduce aliasing effects.

Next to direction, identity, type of motion, speed, and type of stimuli, a further mode of variation will consist in the addition of noise, and distortion through scrambling of the image contents.

2.4. Stimuli Database

The database is available to DIRAC members and currently consists of (state Jan. 1, 2007):

- MoCap data: 36 sequences, 10s each, available as `*.c3d`, `*.V` and `*.csv` files
- corresponding subject-specific skeleton data: `*.vsk` files
- Visual stimuli [point-light]: 72,600 frames at 60 Hz each, `*.PNG` image format
- Visual stimuli [stick figure]: 72,600 frames at 60 Hz each, `*.PNG` image format
- Visual stimuli [humanoid]: 72,600 frames at 60 Hz each, `*.PNG` image format

Total database size (compressed) is approx. 2 GB.

3. Conclusion

In this deliverable, we have presented the basic experimental setup and stimuli used in the first round of neurophysiological experiments at KUL, as well as a new set of stimuli that has been developed in collaboration between ETHZ and KUL. This new data set will be the basis for a second round of neurophysiological, as well as of computational experiments on action recognition.

In particular, KUL will use the complex action stimuli developed in collaboration with ETHZ in a subsequent study in which KUL will teach the animals to discriminate the direction of running/walking and this for different levels of dynamic noise masking the stimuli. The noise manipulation will allow us to measure psychometric functions of biological motion discrimination in the monkeys and subsequent single-cell recordings will measure the selectivity of STS neurons for the same displays when the animals are performing the discrimination. Thus, we can relate behavioral discrimination of biological motion with discrimination capacity of single neurons for the same stimuli.

In parallel, ETHZ will perform a statistical analysis of the motion-capture data from which those new stimuli were created in order to determine the statistical dependencies and modes of variation of the data. The analysis will focus on static and dynamic aspects of body motion, as well as on the appearance of moving bodies in monocular video. This will yield statistical models of biologically plausible body postures and motion patterns which will be used both to support the neurophysiological experiments and to build computational action recognition algorithms. The ETHZ action recognition approach will be gradually extended with different ways to integrate detection and tracking elements in order to reduce pose ambiguity and obtain more reliable results. In particular, the statistical model learned from MoCap data will be used to constrain body poses, as well as their succession.