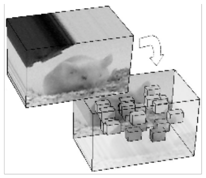
Behavior Recognition in Video With Extended Models of Feature Velocity Dynamics

Ross Messing and Chris Pal

Spring Symposium 2009

Standard Features

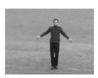


(image from Dollar et al. 2004)

- •Local space-time patch features are clustered, treated as bag-of-words
- •Ignores any information outside the local neighborhood defined by the patch

Standard Datasets

Low resolution videos of simple behaviors
Example: KTH dataset
160 x 120 pixels
6 behaviors characterized by broad cyclic motion:
Boxing, Clapping, Waving, Walking, Running, and Jogging



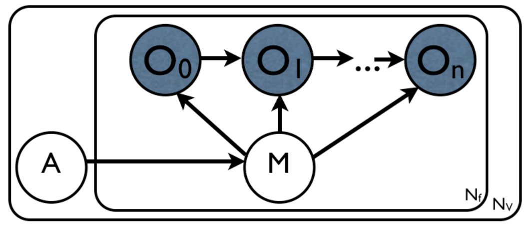
Our Dataset

1280 x 720 pixels
10 activities of daily living:
Dial Phone, Answer Phone, Drink Water, Eat Snack Chips, Lookup in Phonebook, Write on Whiteboard, Peel Banana, Chop Banana, Eat Banana, Use Silverware

Feature Velocity Dynamics

•KLT Tracker gives us velocity histories of tracked keypoints

•Each feature's velocity history is an observation sequence

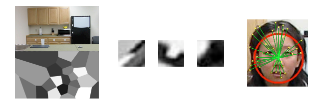


$$
\begin{aligned}
P(A,O) &= \sum_M P(A,M,O) = \\
&P(A)\prod_f^{N_f}\sum_i^{N_m} P(M_f^i|A)P(O_{0,f}|M_f^i) \\
&\prod_{t=1}^{T_f} P(O_{t,f}|O_{t-1,f},M_f^i)
\end{aligned}
$$

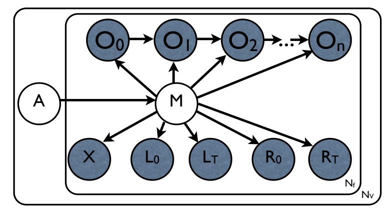
Model	Performance on KTH	Performance on new dataset
Temporal Templates (Bobick and Davis 2001)	X	33%
Cuboids (Dollar et al. 2004)	66%	36%
Space-Time Interest Points (Laptev et al. 2008)	80%	59%
Feature Velocity Dynamics (Us)	74%	63%

Augmentations

•Feature Velocity Dynamics directly represent the motion of a single point.
•How can we add information to this feature that will increase its descriptive power?
•What information would we want to add?

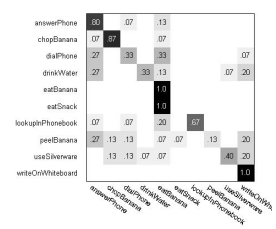


•All augmentations are codebooked
•Each mixture component has a distribution over each augmentation



$$
\begin{aligned}
&P(A,L_0,L_T,R_0,R_T,X,O) = \\
&\quad \sum_M P(A,M,L_0,L_T,R_0,R_T,X,O) = \\
&P(A)\prod_f^{N_f}\sum_i^{N_m} P(M_f^i|A)P(L_{0,f}|M_f^i)P(L_{T,f}|M_f^i) \\
&P(R_{0,f}|M_f^i)P(R_{T,f}|M_f^i)P(X_f|M_f^i) \\
&P(O_{0,f}|M_f^i)\prod_{t=1}^{T} P(O_{t,f}|O_{t-1,f},M_f^i)
\end{aligned}
$$

STIP features (Laptev et al. 2008)
55%



Our method with augmentations
89%