TOWARDS DEFINING GROUPS AND CROWDS IN VIDEO USING THE ATOMIC GROUP ACTIONS DATASET

Ricky J. Sethi

Fitchburg State University, Computer Science Department

ABSTRACT

Understanding group activities is an essential step towards studying complex crowd behaviours in video. However, such research is often hampered by the lack of a formal definition of a group, as well as a dearth of datasets that concentrate specifically on Atomic Group Actions. [1] In this paper, we provide a quantitative definition of a group based on the Group Transition Ratio (G_{tr}); the G_{tr} helps determine when individuals transition to becoming a group (where the individuals can still be tracked) or a crowd (where tracking of individuals is lost).

In addition, we introduce the Atomic Group Actions Dataset, a set of 200 videos that concentrate on the atomic group actions of objects in video, namely the group-group actions of *formation*, *dispersal*, and *movement* of a group, as well as the group-person actions of *person joining* and *person leaving* a group. We further incorporate a structured, end-to-end analysis methodology, based on workflows, to easily and automatically allow for standardized testing of new group action models against this dataset.

We demonstrate the efficacy of the G_{tr} on the Atomic Group Actions Dataset and make the full dataset (the videos, along with their associated tracks and ground truth, and the exported workflows) publicly available to the research community for free use and extension at at `http://research.sethi.org/ricky/datasets/`.

***Index Terms*—** Group Action Detection, Group Action Dataset, Atomic Group Actions

1. INTRODUCTION

Crowd analysis is a growing area of study in computer vision. However, in order to examine crowd activities, it is necessary to first understand group activities [1, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Group activities, in turn, can be thought of as being composed of Atomic Group Actions. Following the formulation of [1, 2, 3], we use the term **Atomic Group Action** to refer to simple, uniform motion patterns involving three or more, possibly interacting, objects in video, typically lasting for short durations of time. In contrast, the term **Group Activity** refers to a complex sequence of Atomic Group Actions performed by a group of (three or more) possibly interacting objects, typically characterized by much longer temporal durations.

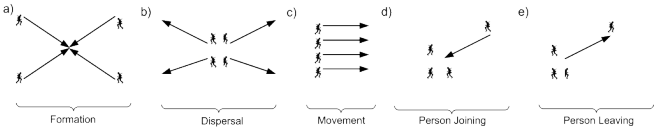


Fig. 1. Atomic Group Actions: group-group actions are shown in a) *formation*, b) *dispersal*, and c) *movement*, as well as group-person actions in a d) *person joining* and e) *person leaving* a group.

Although, as [2] notes, there is no hard boundary between an action and an activity, most researchers [1, 4, 5, 6, 7, 8, 9, 10] refer to three basic motions as being fundamental motions for group-group interaction and two for group-individual interaction. On this basis, we also define Atomic Group Actions to be three or more objects exhibiting the group-group actions of *formation*, *dispersal*, and *movement*, as well as group-person actions of *person joining* and *person leaving* a group, as seen in Figure 1.

2. DEFINING A GROUP

Thus, an analysis of Atomic Group Actions can lead to better insight into complex group and crowd activities and their analysis. However, a significant problem arises in regards to defining a group: although most people intuitively know what a group is, there is no formal definition of a group in computer vision to the best of our knowledge. One of the main contributions of this paper is to use a physics-inspired methodology for modelling the transition of Individuals $\Rightarrow$ Groups $\Rightarrow$ Crowds analogous to the transition of Individual Particles $\Rightarrow$N-Body $\Rightarrow$Fluids in fluid dynamics.

In fluid dynamics, the Knudsen Number determines the transitions from individual particles to fluids in physics, we model the transition from individuals to groups/crowds using a transition number similar to the Knudsen Number [13]. The

[1] Here, we distinguish between the atomic motion of individual objects and the atomic motion of groups of objects. The term action in Atomic Group Action means an atomic interaction movement of three or more objects in video; a group activity, then, is composed of multiple actions by a group or multiple groups of interacting objects [1, 2, 3].

Knudsen number is a simple ratio between the mean free path of particles, λ, and a representative physical scale called the characteristic length, L: $K = \frac{\lambda}{L}$. The intuition behind this ratio is as follows: if the mean free path (the average distance a particle goes before it encounters another particle) is much smaller than some characteristic length (e.g., an opening the particles have to travel through), then the particles will go through in groups, behaving as a fluid. If the mean free path is much larger than that same characteristic length, however, then the particles would likely go through individually, thus behaving as individual particles.

Using similar reasoning, we thus define the *Group Transition Ratio,* G_{tr}, as:

$$G_{tr} = \frac{L}{\lambda} \tag{1}$$

where λ is the mean free path and L is the characteristic length. In fact, in fluid dynamics, there is no formal, exact approach to determining the characteristic length, L. Instead, the characteristic length is usually a convenient reference length that is a constant of a given configuration. For video, we could use some reference length in the video as the characteristic length if we know it a priori, as in the videos in this dataset; but, if not, we use the mean relative distance between groups as the characteristic length, analogous to a normalization constant.

Thus, Equation (1) is actually the inverse of the Knudsen Number since we use the average relative distance between all objects as the characteristic length, L. This is because a group/crowd will have a very small average relative distance between the individuals (the individuals are bundled close together); however, a dispersed set of individuals would have a much larger average relative distance between its individuals, especially compared to the mean free path, which depends on the width/volume of an object in the video. When L is a representative physical length scale in the video (even when that varies with time), then it should be bigger than the mean free path for a group and we can use the same Knudsen Number as in fluid dynamics, instead of its inverse. The mean free path is given as:

$$\lambda = \frac{1}{\sqrt{2}\pi\mu\rho} \tag{2}$$

where μ is the mean width of an object and ρ is the number density of objects; for images, we use the area in ρ but, if 3-d information of objects is available, we utilize volume in ρ and use μ^2, as in the Knudsen Number.

Interpretation: In fluid dynamics, once this ratio becomes small, the particles make a transition from individual particles to a fluid. The intermediate region is the transition region when the collection is neither individual particles nor a fluid. This is the region we associate with groups, intermediate between individuals and crowds. Thus, we utilize these same ideas for analysis of multi-object activities and, in a similar fashion, when $G_{tr} \ll 1$ for a collection of objects, we label them as a crowd; when $G_{tr} \gg 1$, we label them as individual objects; finally, when $G_{tr} \sim 1$ (empirically between 0.1 and 1.5) and they're in a transition region, we label them as a group. In this way, we are able to quantify the ideas of groups and crowds as follows:

$$\begin{cases} G_{tr} \lll 1 & \text{Crowd} \\ G_{tr} \sim 1 & \text{Group} \\ G_{tr} \ggg 1 & \text{Individuals} \end{cases} \tag{3}$$

Scale Invariance: This analysis shows the Group Transition Ratio in Equation (1) is scale invariant since the scale length factor is in both the numerator (L) and the denominator (λ) and would thus cancel out. This is also why simple clustering would not be able to do the same thing as our Group Transition Ratio in categorizing groups/crowds and would not work in both high- and low-resolution domains, as our methodology is able to do. However, spatial clustering could still be useful in sub-group identification.

Advantages: One of the main advantages of this approach is the robustness to errors in generation of object positions: short-term errors in positional data would not affect the performance of this method since we fit to a model. Fitting to a model allows our method to work without significant training data or classifiers. One of the main contributions of this method is to identify when a scene becomes too cluttered, and hence tracking becomes difficult, and then treat the entire collection of objects as a single group. Thus, we are able to work with realistic trackers. Finally, the formalism afforded by our method provides a framework that is extensible with more complex models to an even wider variety of situations and domains. This method can also be used to search for activities given a single query video, since it is unrealistic to assume that multiple examples will be available for highly complex activity classes. Once a crowd has been identified as being formed, any pre-existing crowd analysis approaches can be employed to examine the crowd's dynamics.

2.1. Limits of Tracking

Based on this model, we can search for activities involving multiple objects and analyze group formations and interactions. This is especially helpful when it is necessary to track a collection of individuals that are fast forming a group. When a large number of targets are in close proximity, tracking each individual target becomes very difficult; in such cases, it may be more desirable to treat them as a group or a crowd, which can then be tracked as a single entity.

The Group Transition Ratio (G_{tr}) can identify exactly such situations where individual targets can no longer be tracked. We identify this as formation of a group and then a crowd. The group/crowd can then by tracked as a single entity.

Almost all current methods of crowd analysis [14, 2] assume that it is known that the scene consists of a crowd. Often, we have situations where individuals merge together to form a group and cannot be tracked separately any more. This detection of transitions from individuals to groups and crowds can thus help identify when to stop tracking individuals and start tracking the entire group as a single entity. Once such transitions are identified, pre-existing group/crowd analysis approaches (such as [15, 16, 17]) can be employed to examine the group/crowd's dynamics.

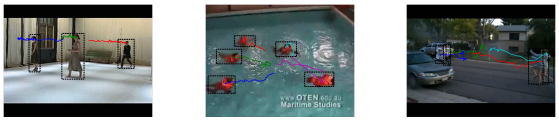
3. ATOMIC GROUP ACTIONS DATASET

In addition to the lack of a formal definition of groups, there is a lack of datasets for evaluating Atomic Group Actions. Most researchers rely on either creating a custom dataset, using parts of pre-existing datasets for the subset of activities they want to examine [1, 18, 4, 6, 9, 19, 20, 10, 8, 21], or utilizing simulations and synthetic agents [22, 21]. In fact, some existing datasets confuse simple activities with atomic actions, as simple (or complex) group activities can often be de-composed into their atomic components. Most current datasets are thus plagued by several limitations, from confusing Atomic Group Actions with Group or Crowd Activities, to using ad hoc components of other datasets, to simply creating a dataset or simulations to demonstrate their specific method only. The Atomic Group Actions Dataset aims to overcome these limitations and provide a dataset that concentrates solely on Atomic Group Actions but is not targetted to any specific method.

The Atomic Group Actions Dataset was collected by recruiting people via Amazon's Mechanical Turk (MTurk) to upload or find public domain videos showing Atomic Group Actions. Almost all the videos were obtained from a combination of YouTube or public domain datasets. There are three categories of videos showing group-group Atomic Group Actions (*formation*, *dispersal*, and *movement*) and two categories of videos showing group-person Atomic Group Actions (*person joining* and *person leaving*), as shown in Figure 1. There are 40 videos in each category for a total of 200 videos.

The videos were altered to ensure uniformity and all videos have the following characteristics: Cropped to 640 x 480 pixels; Frame rate of 30fps; File that includes the origin URL for provenance; Image that shows the Ground Truth for that video with bounding boxes around objects of interest, as well as an outline of their approximate trajectory; Metadata file that includes a ground truth classification, the filename, and the tracks for the objects of interest.

Sample clips for each category can be seen in Figure 2. Each video is a few seconds in length, the shortest is 1 sec and the longest is 15 secs, with most averaging about 5 seconds. For each video sequence, we manually cropped the video, converted it as per the characteristics above, and assigned a ground truth Atomic Group Action category.



(a) Sample frames of the *group formation* Atomic Group Action. Here, we see three sample clips from the *group formation* category, with bounding boxes around the objects and their trajectories overlaid.

(b) Sample frames of the *group dispersal* Atomic Group Action. Here, we see three sample clips from the *group dispersal* category, with bounding boxes around the objects and their trajectories overlaid.

Fig. 2. Sample frames from video clips for all three Atomic Group Action categories: *group formation* and *group dispersal* dispersal. For want of space, frames from *group movement, person joining* and *person leaving* group actions are not shown here.

We used a tracker based on the particle filter algorithm [2, 23] with manual initializations in order to track each objects' motion trajectory. These tracks are included in the metadata file that accompanies each video. In our implementation, we could not track fast motion very well; because of that, the track results became a little bit shorter and resulted in lower results for our models based on relative distance (e.g., for the phase space model this resulted in a lot of noise in training/testing and led to lower accuracy). In addition, the tracking fails quite often due to occlusions, dropped frames, etc.; thus, manual re-initializations were used to reacquire the tracks. Some clips had perspective motion in the z-direction (into or out of the screen) that led to tracking errors, as well. In future work, we will consider a motion segmentation algorithm with automatic re-initializations for automatic tracking [23].

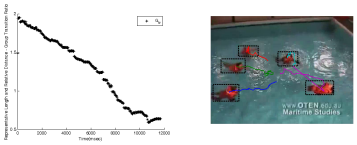
Table 1. Cumulative Statistical Comparison of Group-Group **(G-G)** and **All** atomic group actions. In (a), we see the cumulative statistics for both Group-Group and All atomic group actions. In (b), we see the Confusion Matrix for the Group-Group atomic group actions.

(a)

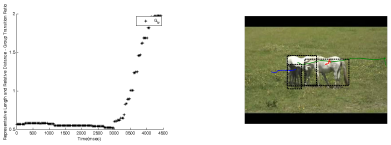
	G-G	All
F-Meas.	0.84	0.60
EER	0.08	0.10
MAP	0.82	0.67
Prec.	0.84	0.60
Recall	0.84	0.60

(b)

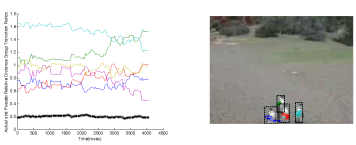
	For.	Dis.	Mov.
For.	1	0	0
Dis.	0	0.8	0.2
Mov.	0.2	0.2	0.6



(a) *Group Formation*: G_{tr} decreases from about 2 to about 0.5. Sample image shows swimmers converging upon a central object, forming a group.

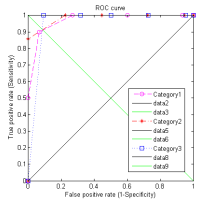


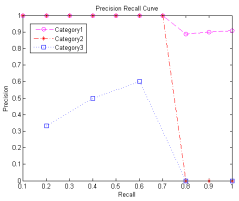
(b) *Group Dispersal*: G_{tr} increases from about 0.5 to about 2.0. Sample image shows a group of horses dispersing when startled.



(c) *Group Movement*: G_{tr} stays steady at about 0.2. Sample image shows penguins waddling along together.

Fig. 3. In (a) - (c), we show the G_{tr} vs. time graph and a sample frame for all three group-group Atomic Group Actions: *group formation*, *group dispersal*, and *group movement*. They show how the G_{tr} decreases with time for *group formation* in (a), increases with time for *group dispersal* in (b), and stays the same for *group movement* in (c). For want of space, group-person Atomic Group Actions of *person joining* and *person leaving* are not shown here but will be released.





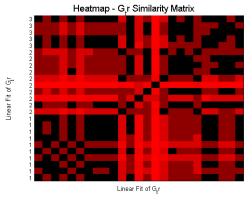


Fig. 4. ROC curve, PRC, and Heatmap for the group-group actions: Category 1 (*group formation*), Category 2 (*group dispersal*), and Category 3 (*group movement*). They show that Cat 3 was hardest to classify while Cat 1 was the easiest. Please note: Cat 3 has the smaller square in the heatmap as group-person action videos were added to Cat 1 and 2, doubling their size.

4. EXPERIMENTS

We use the G_{tr} to identify when collection of objects is individuals, groups, or crowds. Then, we use the time variance of the G_{tr} to determine when a collection of objects transitions between being individuals, groups, or crowds. In this paper, we implemented the G_{tr} in matlab as a component in the Wings workflows system and tested it against the Atomic Group Actions dataset. We first show some qualitative results showing the G_{tr} in action. In Figure (3), we show how the G_{tr} varies with time to characterize the action in a video; here, we show results for all five categories of the G_{tr} graph along with representative frames from the video overlaid to show the action. In order to characterize group-group atomic group actions, we only utilize the G_{tr} by doing a linear fit on the G_{tr} vs time graph. The thresholds were empirically determined to be $G_{tr} \lesssim 0.1$ for crowds, $0.1 \lesssim G_{tr} \lesssim 1.5$ for groups, and $G_{tr} \gtrsim 1.5$ for individuals.

However, for group-person atomic group actions, we use a two-step process where we first utilize the G_{tr} and then confirm it via a linear fit on the trajectories vs time graph of the objects followed by a K-means clustering to see if a single trajectory veers off from the cluster of other trajectories (*person leaving*) or veers towards the cluster of other trajectories (*person joining*). In future work, we will incorporate more complex models for group-person actions but the main focus of this work is on analysis of the group-group actions via the G_{tr} model.

In Figure (3) (a) - (c), we show the G_{tr} vs. time graph and a sample frame for all three group-group Atomic Group Actions: *group formation*, *group dispersal*, and *group movement*. They show how the G_{tr} decreases with time for *group formation* in (a), increases with time for *group dispersal* in (b), and stays the same for *group movement* in (c).

We also show some quantitative results using standard statistical evaluation. In Figure 4, we show the ROC curve, PRC, and Heatmap for the group-group actions: Category 1 (*group formation*), Category 2 (*group dispersal*), and Category 3 (*group movement*). They show that Category 3 was hardest to classify while Category 1 was the easiest. Please note: Category 3 has the smaller square in the heatmap as group-person action videos were added to Category 1 and 2, doubling their size. We also show the Confusion Matrices, EER, and F-Measure for both group-group and group-person atomic group actions in Table 1. For the overall table of confusion for *all* atomic group actions, we had TP=0.12, FN=0.08, FP=0.08, and TN=0.72. As mentioned earlier, our main purpose in this paper is to quantify group-group atomic group actions using the G_{tr} and we use a simple baseline method based on trajectories for group-person classifications which results in lower classifications. Other issues affecting accuracy include videos with small number of objects for clustering (three or less lead to lower results), sufficient number of videos used for training (dependent on the method utilized), and perspective issues with motion in the z-direction (to be improved in future tracker implementations).

5. REFERENCES

[1] Tian Lan, Yang Wang, Weilong Yang, Stephen Robinovitch, and Greg Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[2] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *CSVT*, 2008.

[3] A.R. Ahad, *Computer Vision and Action Recognition: A Guide for Image Processing and Computer Vision Community for Action Understanding*, Atlantis Ambient and Pervasive Intelligence. Atlantis Press, 2011.

[4] Tian Lan, Yang Wang, Greg Mori, and Stephen Robinovitch, "Retrieving actions in group contexts," in *International Workshop on Sign Gesture Activity*, 2010.

[5] Stefano Pellegrini, Andreas Ess, and Luc Van Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in *ECCV*, Berlin, Heidelberg, 2010, pp. 452–465.

[6] Dong Zhang, Daniel Gatica-Perez, and Samy Bengio, "Modeling individual and group actions in meetings with layered HMMs," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 509–520, 2006.

[7] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?," in *CVPR*, Washington, DC, USA, 2011, pp. 1345–1352.

[8] R.J. Sethi and A.K. Roy-Chowdhury, "Physics-based activity modelling in phase space," *ICVGIP*, 2010.

[9] Yue Zhou, Bingbing Ni, Shuicheng Yan, and Thomas S. Huang, "Recognizing pair-activities by causality analysis," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 1, pp. 1–20, 2011.

[10] Utkarsh Gaur, Bi Song, and Amit K Roy-Chowdhury, "Query-based Retrieval of Complex Activities using "Strings of Motion-Words"," in *WMVC*, 2009.

[11] Weiming Hu, Tieniu Tan, Liang Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 34, no. 3, pp. 334–352, Aug 2004.

[12] Beibei Zhan, Dorothy N. Monekosso, Paolo Remagnino, Sergio A. Velastin, and Li-Qun Xu, "Crowd analysis: A survey," *Mach. Vision Appl.*, vol. 19, no. 5-6, pp. 345–357, Sept. 2008.

[13] C. Shen, *Rarefied Gas Dynamics: Fundamentals, Simulations and Micro Flows (Heat and Mass Transfer)*, Springer, 2005.

[14] Julio Cezar Silverira Jacques Junior, Soraia Raupp Musse, and Claudio Rosito Jung, "Crowd analysis using computer vision techniques," *Signal Processing*, , no. September, pp. 66–77, 2010.

[15] Min Hu and Saad Ali, "Detecting global motion patterns in complex videos," in *ICPR*. Dec. 2008, pp. 1–5, Ieee.

[16] Saad Ali and Mubarak Shah, "Floor fields for tracking in high density crowd scenes," in *ECCV*, 2008, pp. 1–14.

[17] Mikel Rodriguez, Saad Ali, and Takeo Kanade, "Tracking in unstructured crowded scenes," in *ICCV*, 2009, number September.

[18] Tian Lan, Yang Wang, Weilong Yang, and Greg Mori, "Beyond actions: Discriminative models for contextual group activities," in *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[19] Yue Zhou, Shuicheng Yan, and Thomas S Huang, "Pair-Activity Classification by Bi-Trajectories Analysis," in *CVPR*, 2008.

[20] Bingbing Ni, Shuicheng Yan, and Ashraf Kassim, "Recognizing Human Group Activities with Localized Causalities," in *CVPR*, 2009, pp. 1470–1477.

[21] J.C. Nascimento, M.A.T. Figueiredo, and J.S. Marques, "Segmentation and classification of human activities," in *BMVC Workshop on Human Activity Recognition and Modelling*, 2005, number September.

[22] Nuria M. Oliver, Barbara Rosario, and Alex P. Pentland, "A Bayesian computer vision system for modeling human interactions," *PAMI*, vol. 22, no. 8, pp. 831–843, 2000.

[23] Nandita M Nayak, Ricky J Sethi, Bi Song, and Amit K Roy-Chowdhury, "Motion Pattern Analysis for Event and Behavior Recognition," in *Visual Analysis of Humans*, T B Moeslund, L Sigal, V Krüger, and A Hilton, Eds., pp. 289–309. Springer-Verlag, 2011.