

Comparison of descriptors between pilot data and Alon's experiment (V0)

Pilot Analysis - 14 May, 2019

Overview

This document describes the steps to analyse the descriptors between Zhao's pilot data (free-recall under fixed SOA) and Alon's experiment (no time constraint).

Materials/Data

- 20 images
- Descriptors from Alon's 640 experiment
- Translated descriptions of each image (4-5 participants) from Alon's experiment - `translatedEXP5`
- Descriptors removed list from Alon's experiment
- Zhao's pilot data descriptors under 67, 133, 256ms

Steps

1. Combine and pre-processed Zhao's pilot data, then calculate the weighted probabilities (by mean confidence)
2. Read Alon's descriptors and original subject descriptors and calculate the probability.
3. Merging the probabilities of Alon's descriptors with Zhao weighted probabilities (with confidence).

Step 1: Combine and pre-processed Zhao's pilot data, then calculate the weighted probabilities (by mean confidence)

The figure below shows a sample of the data after the normalised weighted probabilities were derived.

- **probability** of a word is the total occurrence of the word under an image per SOA.
- **confidence** of a word is the average participants' confidence ratings of that word under an image per SOA.
- $weighted_prob = probability * confidence$
- $norm_weighted_prob = (probability * confidence) / sum(weighted_prob)$ per image per SOA.

```
## Observations: 1,836
## Variables: 9
## $ img          <chr> "im0000009.jpg", "im0000009.jpg", "im0000000...
## $ soa          <int> 67, 67, 67, 67, 67, 67, 67, 67, 67, 67, 67,...
## $ word         <chr> "boat", "water", "black", "blank", "boy", "...
## $ frequency    <dbl> 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ confidence   <dbl> 3.500000, 3.333333, 1.000000, 1.000000, 1.0...
## $ probability  <dbl> 0.08333333, 0.08333333, 0.04166667, 0.04166...
## $ weighted_prob <dbl> 0.29166667, 0.27777778, 0.04166667, 0.04166...
## $ norm_weighted_prob <dbl> 0.11931818, 0.11363636, 0.01704545, 0.01704...
## $ img_id       <int> 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9...
```

Now, I checked the normalised weighted probabilities sum up to 1 for each image per soa.

Total unique combination of image and SOA is

```
print(count(unique(all_df[,.(img, soa)])))
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     69
```

I expect the same if we sum up the normalised weighted probabilities.

```
prob_check_data <- all_df[, sum(norm_weighted_prob), by = .(img, soa)]
print(sum(prob_check_data$V1))
```

```
## [1] 69
```

Table below shows the top words for each image/soa that have the highest normalised weighted probability.

	img	soa	norm_weighted_prob	word
## 1:	im0000003.jpg	67	0.13333333	building
## 2:	im0000003.jpg	133	0.09395973	buildings
## 3:	im0000003.jpg	267	0.08298755	people
## 4:	im0000003.jpg	267	0.08298755	sky
## 5:	im0000009.jpg	67	0.11931818	boat
## 6:	im0000009.jpg	133	0.12893983	water
## 7:	im0000009.jpg	267	0.10723861	river
## 8:	im0000014.jpg	67	0.05853659	crowd
## 9:	im0000014.jpg	67	0.05853659	people
## 10:	im0000014.jpg	133	0.10891089	bull
## 11:	im0000014.jpg	267	0.17283951	cow

Step 2: Read Alon's descriptors and original subject descriptors and calculate the probability

- Source:
 - Alon's filtered descriptors
 - Alon's original translated descriptions from Shinji's (with max 5 translations per image)

In Alon's dataset,

- **frequency** represents the number of occurrence of a word among the 4-5 translated image descriptions.
- **probability** is the normalised probability, i.e. $frequency / \sum(frequency)$ - Sum of this column per image should be equals to 1.

```
## Observations: 983
## Variables: 4
## $ img_id      <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...
## $ word        <chr> "africa", "america", "asia", "atmosphere", "betwee...
## $ frequency    <dbl> 1, 1, 1, 1, 1, 1, 2, 1, 1, 3, 2, 1, 1, 1, 1, 1, 1, ...
## $ probability  <dbl> 0.01298701, 0.01298701, 0.01298701, 0.01298701, 0....
```

Number of unique images:

```
length(unique(alon_img_desc[, img_id]))
```

```
## [1] 23
```

I expect the sum of probabilities per image is 1, in another words, the sum of all probabilities should be equal to the number of unique images.

```
sum(alon_img_desc[, sum(probability), by=img_id]$V1)
```

```
## [1] 23
```

Removed words

The following words are in Alon's descriptors but unable to find in the translated descriptions, and also they are not part of the removed words.

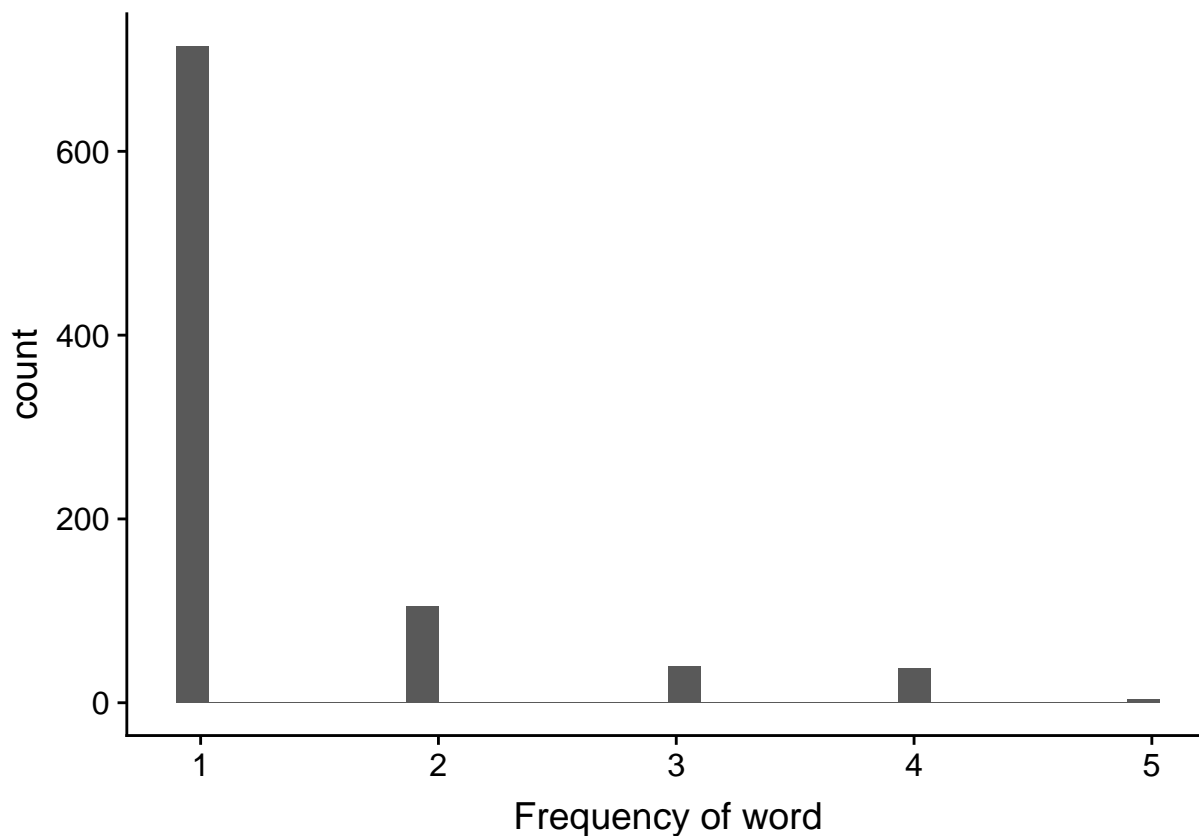
```
##      img_id      word frequency probability
## 1:    2164    smoke           0           0
## 2:    2667  picture           0           0
```

I will remove all words that have zero frequency.

Here is a histogram that shows the number of word frequencies for Alon's descriptors in the images used in pilot.

Most of the descriptors used only once - among 4-5 descriptions. These could reflect the 2-rater words and 4-rater words that Alon used in the experiment.

```
##      frequency total percentage
## 1:           1   714 79.2452830
## 2:           2   105 11.6537181
## 3:           3    40  4.4395117
## 4:           4    38  4.2175361
## 5:           5     4  0.4439512
```



Step 3: Merging the probabilities of Alon's descriptors with Zhao weighted probabilities (with confidence).

In this step, I merged the pilot dataset with Alon's in order to construct the word matrix with different SOAs.

In this data table:

- `p_67` is the probability (not normalised) for a word under SOA 67ms
- `nwp_67` is the normalised weighted probability for a word under SOA 67ms
- `p_133` is the probability (not normalised) for a word under SOA 133ms
- `prob` is the normalised probability from Alon's descriptors (no time limit)

```
## Observations: 2,184
## Variables: 9
## $ img_id <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...
## $ word   <chr> "africa", "america", "asia", "atmosphere", "between", ...
## $ p_67   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ nwp_67 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ p_133  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ nwp_133 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ p_267  <dbl> NA, 0.02272727, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0.02272727...
## $ nwp_267 <dbl> NA, 0.01659751, NA, NA, NA, NA, NA, NA, NA, NA, NA, 0.02489627...
## $ prob   <dbl> 0.01298701, 0.01298701, 0.01298701, 0.01298701, 0.0129...
```

Results

Constructing KL Divergence Matrix

For the pilot data (67, 133, 267ms SOAs) and Alon's (Unlimited), I calculated the *pairwise* KL Divergence (using the normalised probabilities) and constructed a KL matrix.

The actual KL divergence (log2) is calculated using the following command:

```
KLMatrix[row, col] <- KL(rbind(prob[row,], prob[col,]), test.na = FALSE, unit="log2") # measure in bits
```

Note that KL divergence value is not symmetrical, i.e. $KL(P, Q)$ not equals to $KL(Q, P)$.

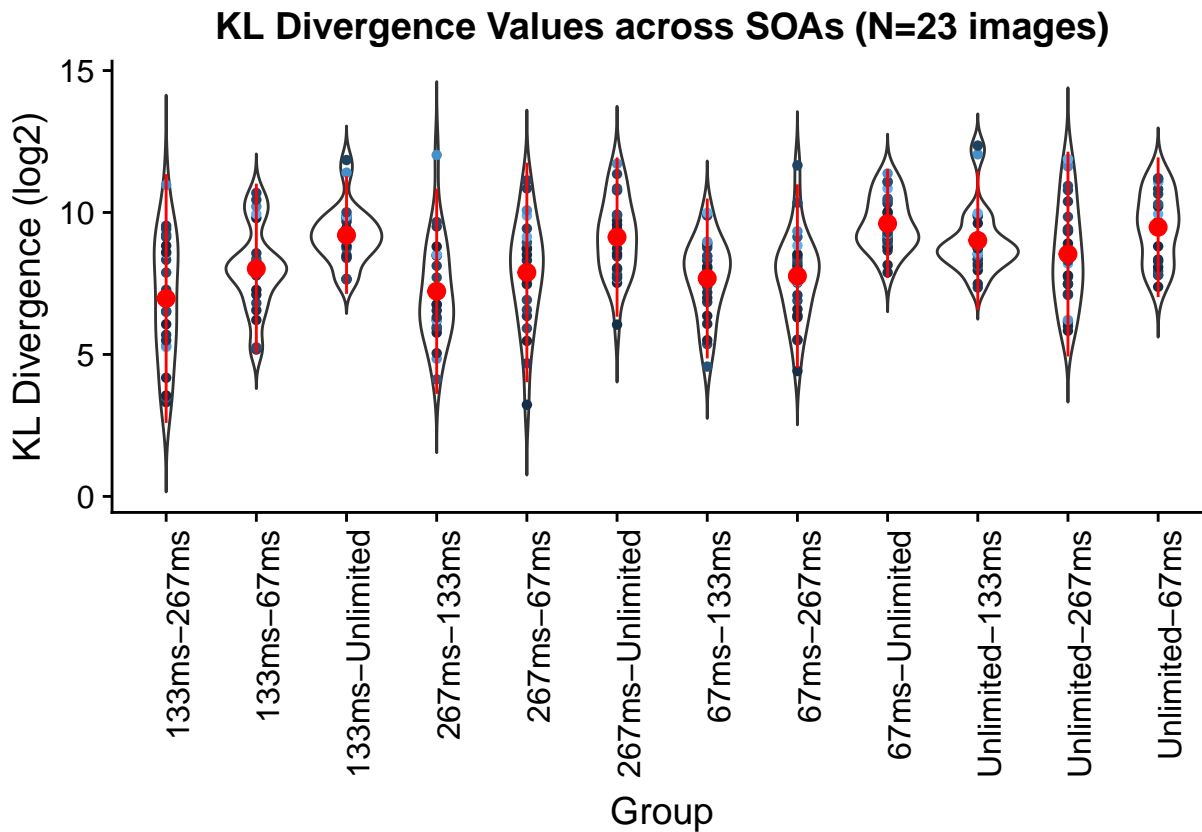
Aggregate the common words between two SOAs

Besides the KL Divergence Matrix, I have aggregated the common-word table that contains the columns:

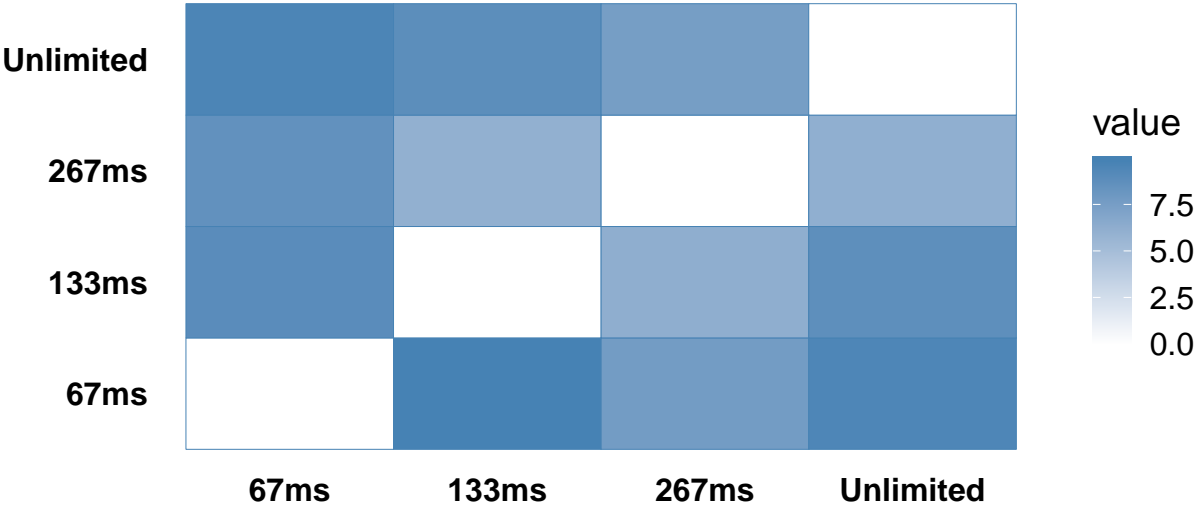
- `all_soas_alon` - These are the words that are common between all SOAs in the pilot and Alon.
- `all_soas` - These are the words that are common between all SOAs in the pilot only.
- `67_alon` - These are the words that are common between 67ms in the pilot and Alon.
- `133_alon` - These are the words that are common between 133ms in the pilot and Alon.
- `267_alon` - These are the words that are common between 267ms in the pilot and Alon.

Summary Statistics

A violin plot showing the mean KL divergence values across 23 images above.

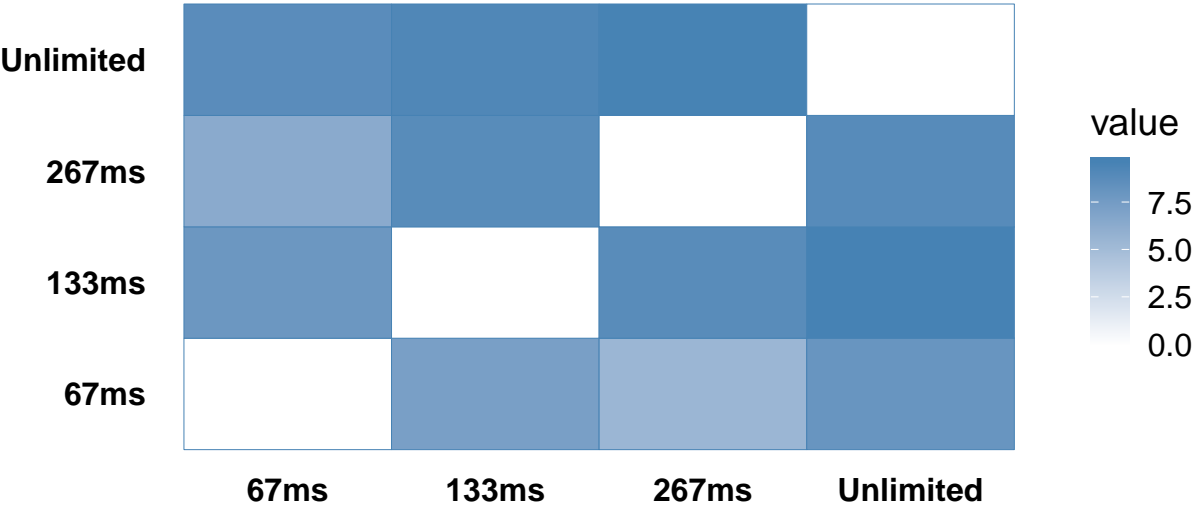


KL Divergence Matrix for Image 3



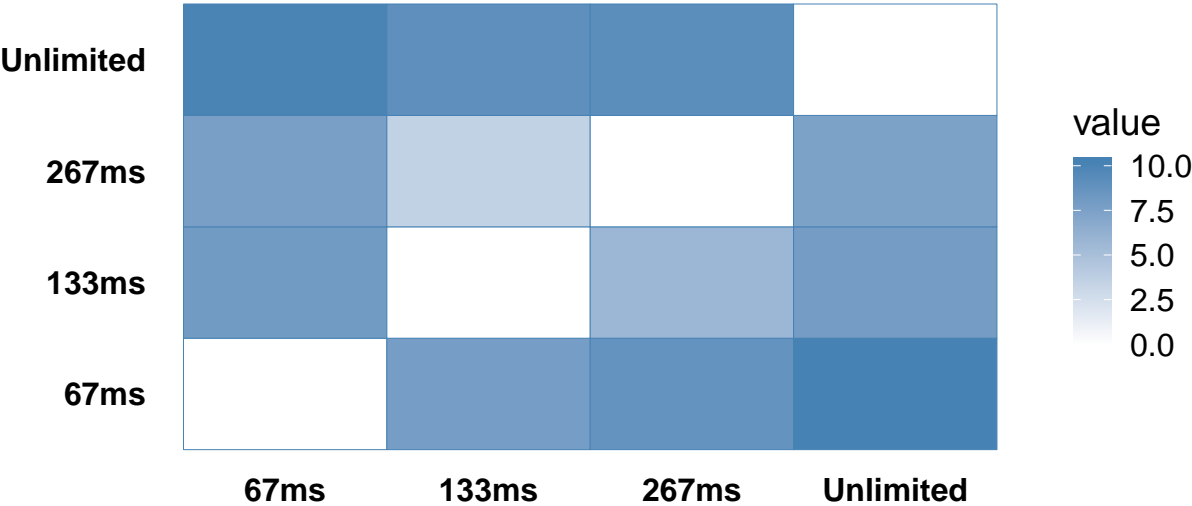
all_soas_alon	all_soas	67_alon	133_alon	267_alon
blue	blue	blue	blue	america
man	man	building	man	blue
road	road	man	people	building
-	tree	road	road	country
-	-	-	sky	man
-	-	-	street	people
-	-	-	-	road
-	-	-	-	sky
-	-	-	-	south
-	-	-	-	street
-	-	-	-	walking

KL Divergence Matrix for Image 9



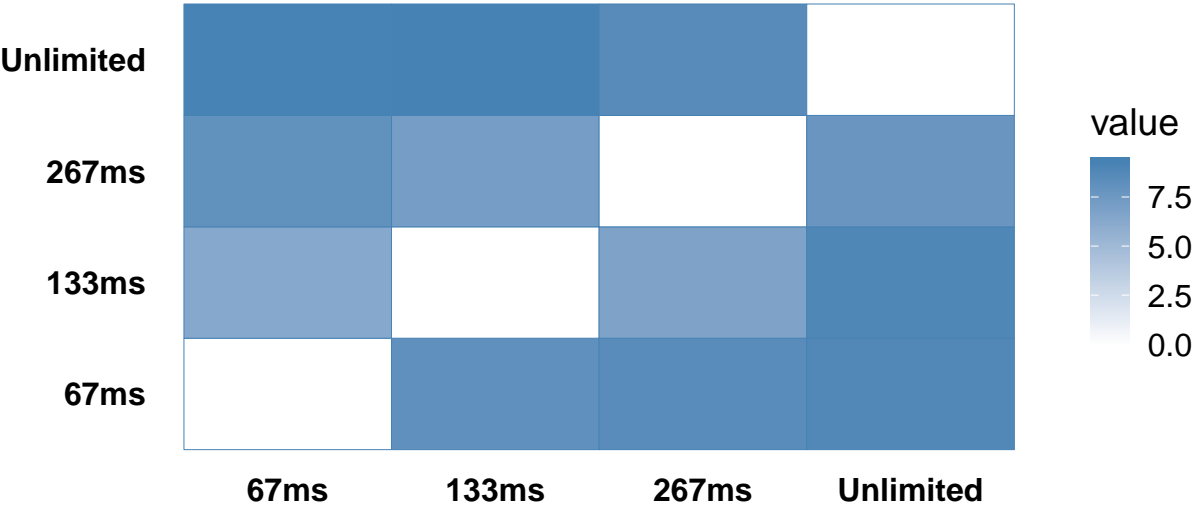
all_soas_alon	all_soas	67_alon	133_alon	267_alon
boat	boat	boat	boat	boat
man	man	canoe	canoe	kayak
-	water	man	good	man
-	-	oar	man	river
-	-	river	oar	-

KL Divergence Matrix for Image 14



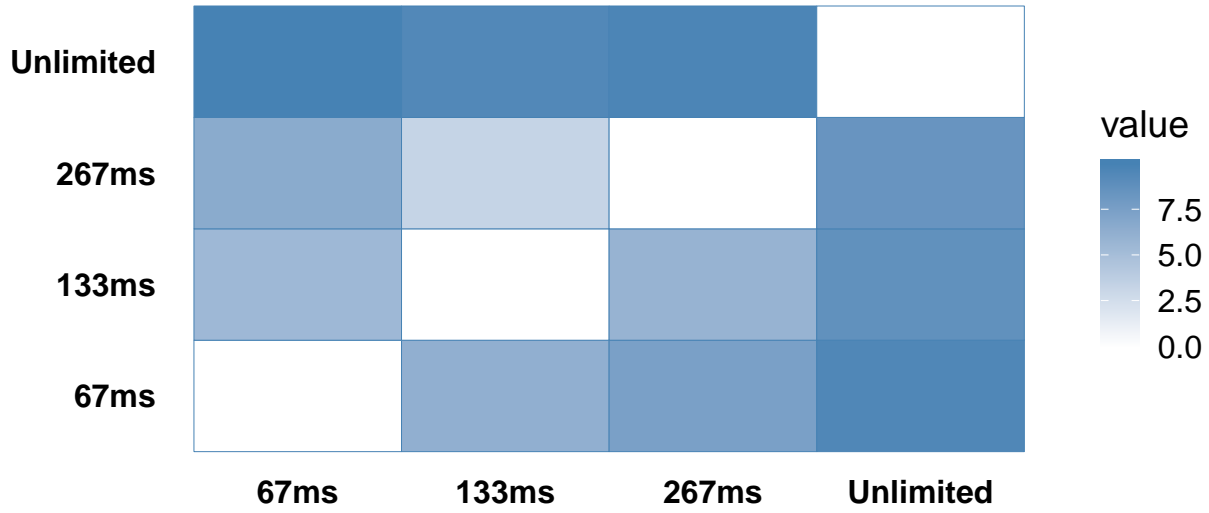
all_soas_alon	all_soas	67_alon	133_alon	267_alon
cow	cow	cow	big	bull
people	crowd	people	bull	cow
-	people	tourist	cow	people
-	road	-	people	statue
-	-	-	statue	-

KL Divergence Matrix for Image 104



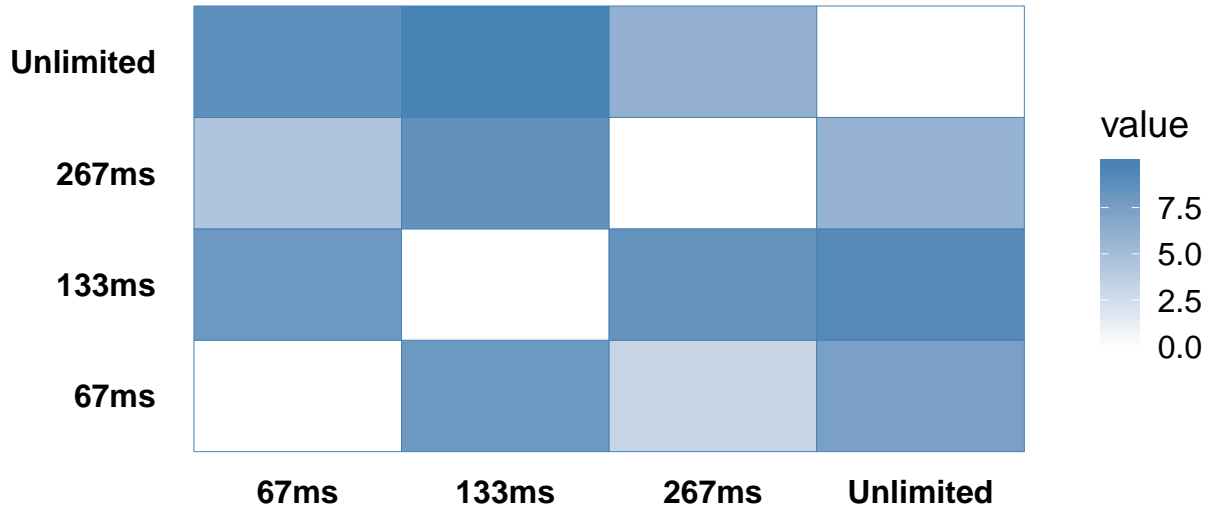
all_soas_alon	all_soas	67_alon	133_alon	267_alon
cow	cow	body	cow	cow
grass	field	cow	goat	goat
-	grass	eating	grass	grass
-	-	grass	-	orange
-	-	sky	-	sky
-	-	-	-	sun

KL Divergence Matrix for Image 277



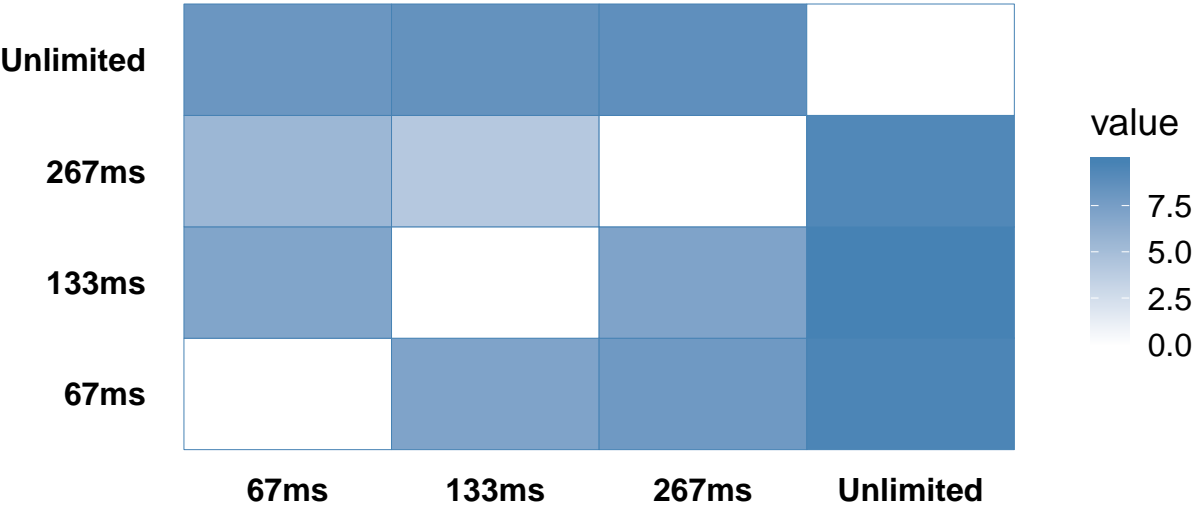
all_soas_alon	all_soas	67_alon	133_alon	267_alon
child	child	baby	baby	boy
woman	grass	child	boy	child
-	mom	woman	child	tree
-	mother	-	good	woman
-	road	-	tree	-
-	woman	-	woman	-

KL Divergence Matrix for Image 321



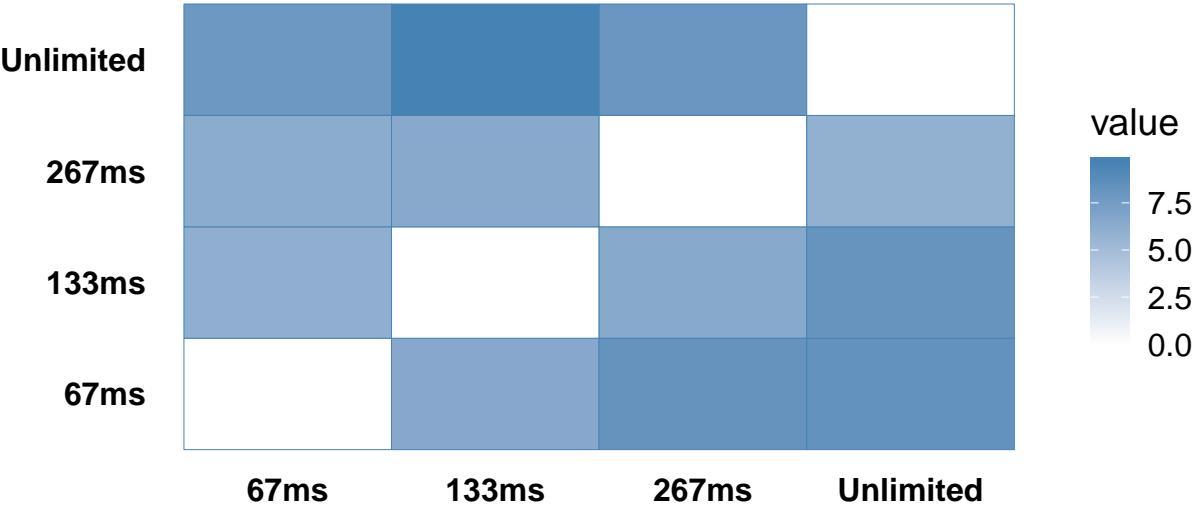
all_soas_alon	all_soas	67_alon	133_alon	267_alon
phone	booth	car	call	black
-	cap	hat	phone	calling
-	men	man	-	car
-	phone	phone	-	hair
-	telephone	talking	-	hat
-	-	-	-	long
-	-	-	-	man
-	-	-	-	phone
-	-	-	-	public
-	-	-	-	talking

KL Divergence Matrix for Image 570



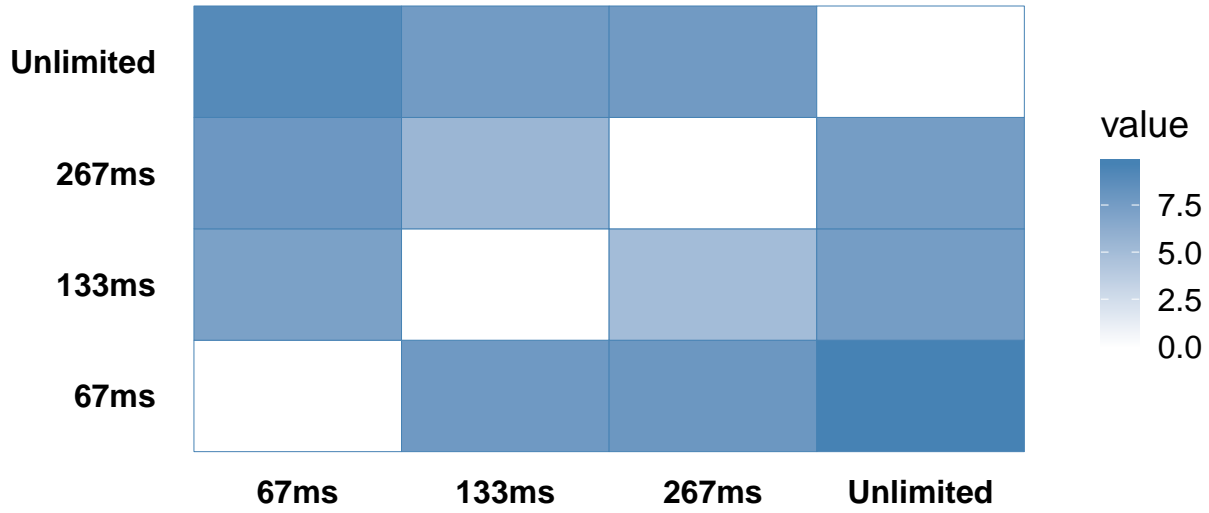
all_soas_alon	all_soas	67_alon	133_alon	267_alon
animal	animal	animal	animal	animal
nature	forest	brown	beautiful	flower
two	grass	eating	brown	nature
-	nature	green	flower	plant
-	tree	nature	nature	two
-	two	two	plant	wild
-	-	-	two	-

KL Divergence Matrix for Image 774



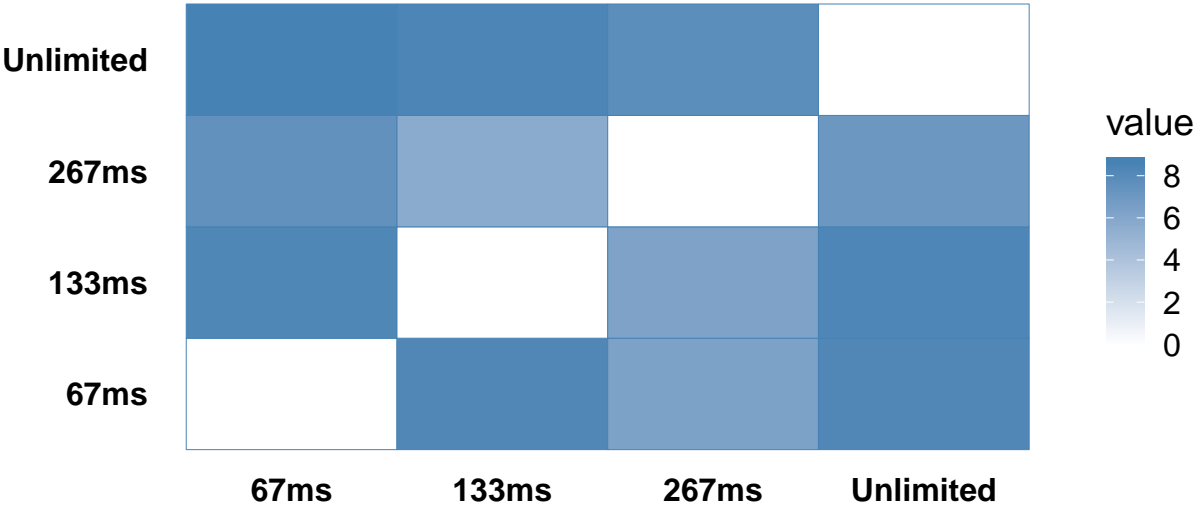
all_soas_alon	all_soas	67_alon	133_alon	267_alon
fire	fire	car	fire	car
people	firemen	engine	people	engine
water	people	fire	red	fire
-	truck	ground	road	hose
-	water	people	water	people
-	-	red	-	water
-	-	road	-	-
-	-	water	-	-
-	-	white	-	-

KL Divergence Matrix for Image 1078



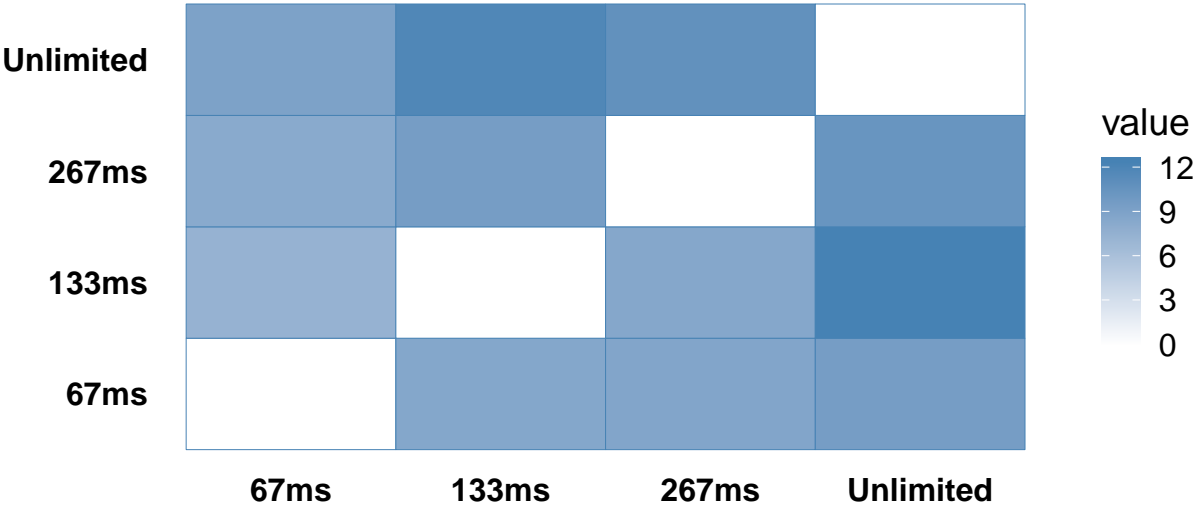
all_soas_alon	all_soas	67_alon	133_alon	267_alon
man	girl	man	grass	bridge
woman	man	talking	man	grass
-	woman	white	people	greenery
-	women	woman	pink	man
-	-	-	shirt	pink
-	-	-	sitting	sitting
-	-	-	talking	tree
-	-	-	tree	woman
-	-	-	woman	-

KL Divergence Matrix for Image 1445



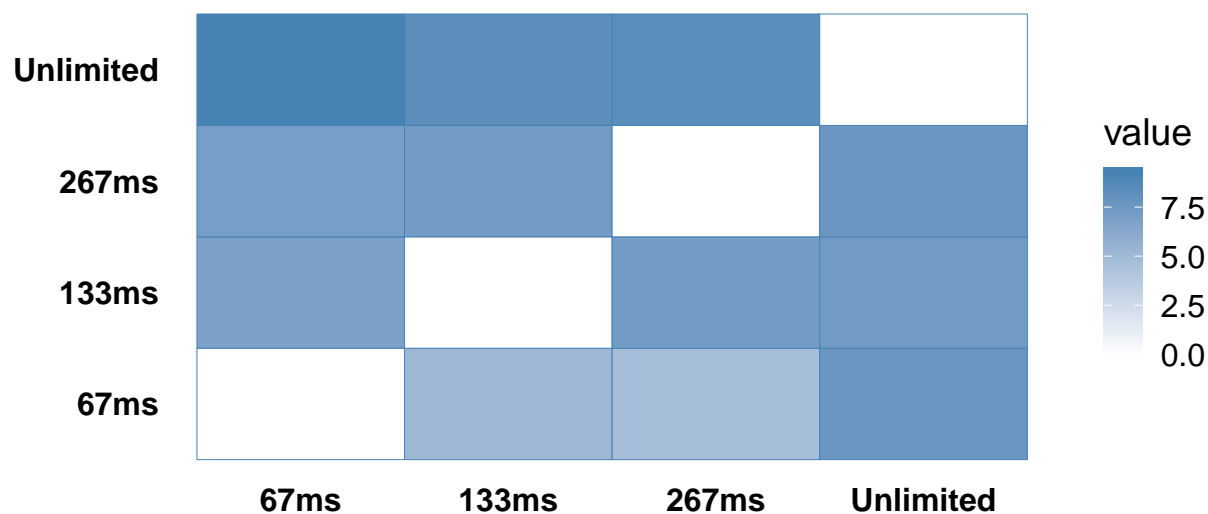
all_soas_alon	all_soas	67_alon	133_alon	267_alon
belt	belt	belt	belt	belt
man	fight	man	black	black
white	man	white	competition	karate
-	shirt	woman	dress	man
-	white	-	fighting	mat
-	-	-	karate	white
-	-	-	man	woman
-	-	-	white	-

KL Divergence Matrix for Image 1684



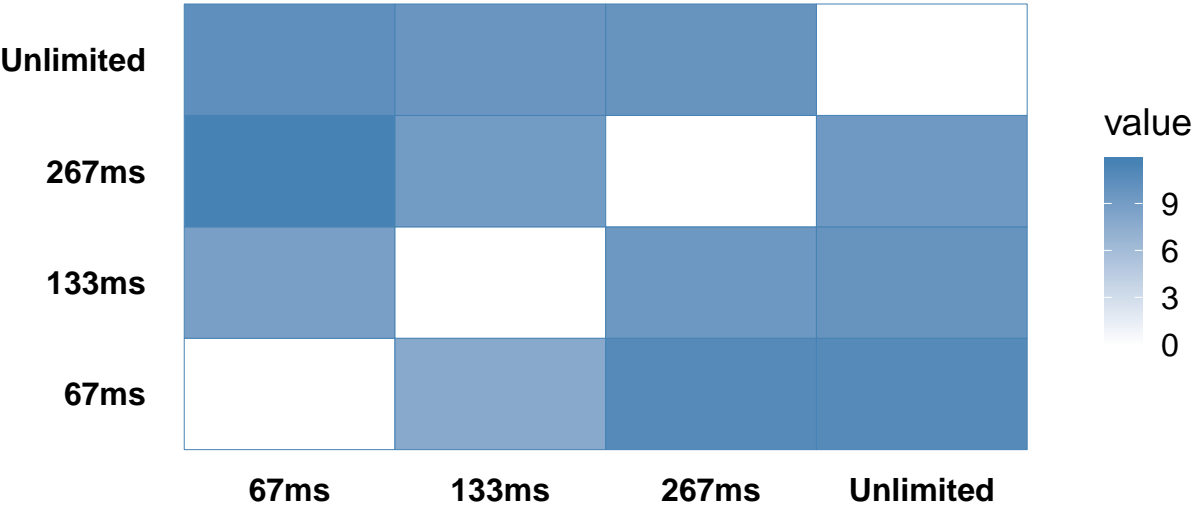
all_soas_alon	all_soas	67_alon	133_alon	267_alon
-	bald	boy	-	ground
-	boat	ground	-	people
-	man	sitting	-	wood
-	sand	wood	-	-
-	tree	xylophone	-	-

KL Divergence Matrix for Image 1715



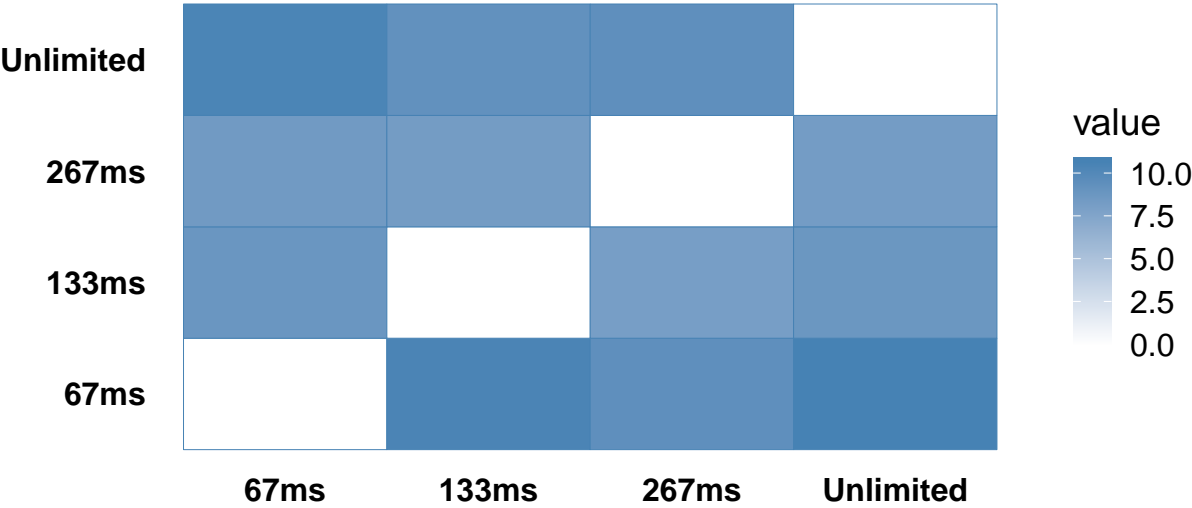
all_soas_alon	all_soas	67_alon	133_alon	267_alon
man	glass	man	man	dust
mask	man	mask	mask	helmet
table	mask	table	table	iron
-	stick	-	white	man
-	table	-	-	mask
-	-	-	-	shovel
-	-	-	-	table

KL Divergence Matrix for Image 1758



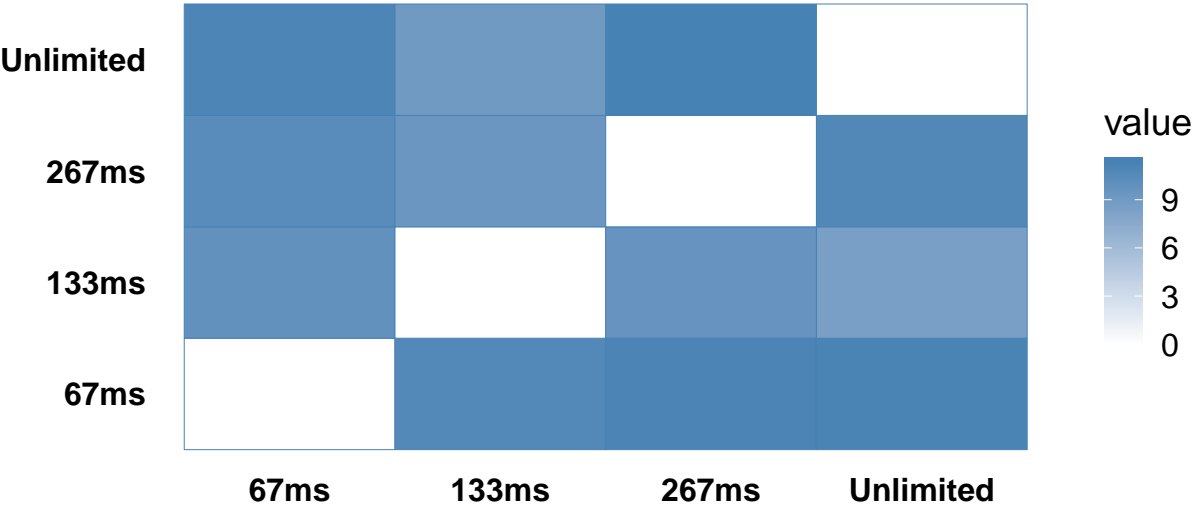
all_soas_alon	all_soas	67_alon	133_alon	267_alon
-	sky	people	canyon	man
-	women	woman	grand	mountain
-	-	-	mountain	-
-	-	-	people	-
-	-	-	seeing	-
-	-	-	shirt	-
-	-	-	top	-

KL Divergence Matrix for Image 2033



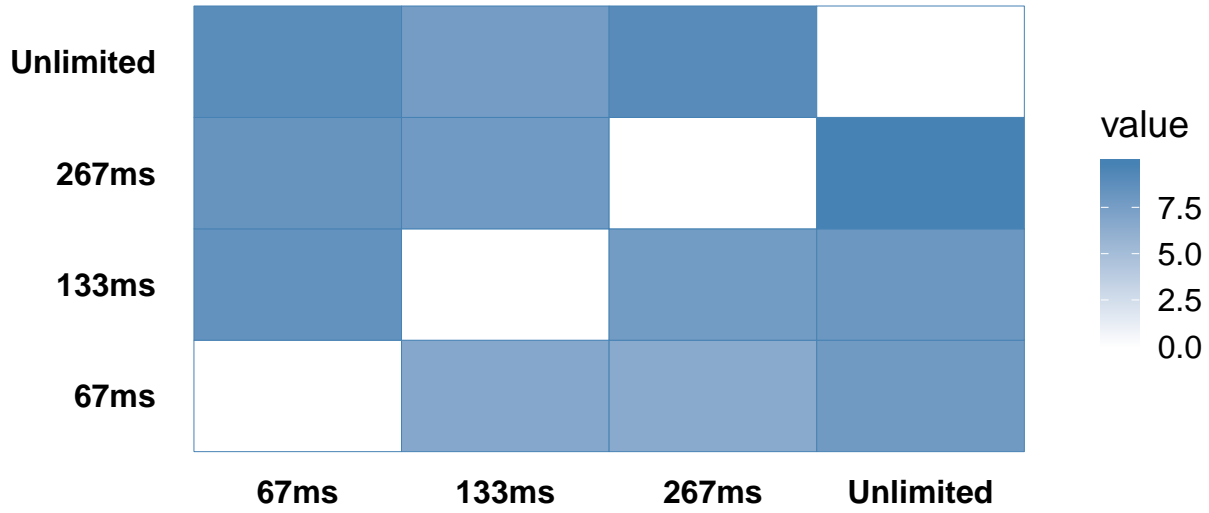
all_soas_alon	all_soas	67_alon	133_alon	267_alon
water	water	man	different	boy
-	-	water	fountain	fountain
-	-	-	park	man
-	-	-	water	water

KL Divergence Matrix for Image 2164



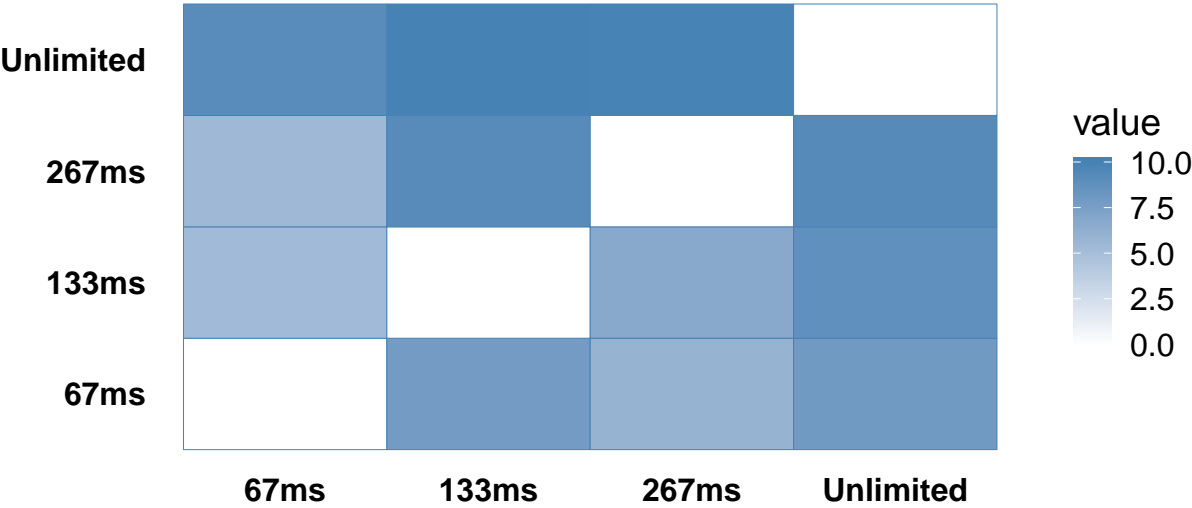
all_soas_alon	all_soas	67_alon	133_alon	267_alon
room	room	people	aisle	room
-	-	room	door	-
-	-	-	factory	-
-	-	-	room	-
-	-	-	wall	-

KL Divergence Matrix for Image 2537



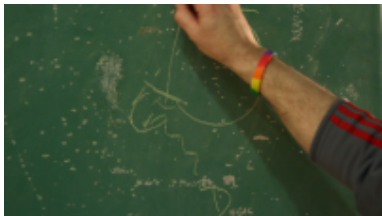
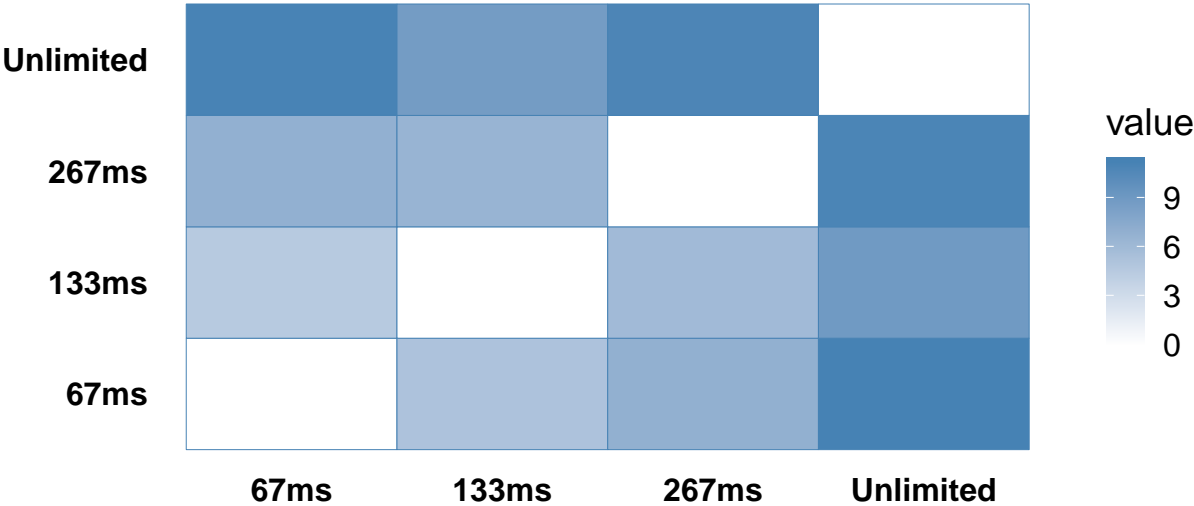
all_soas_alon	all_soas	67_alon	133_alon	267_alon
man	dress	book	book	man
person	man	man	man	person
white	person	people	people	three
-	white	person	person	white
-	-	sitting	show	-
-	-	white	space	-
-	-	-	three	-
-	-	-	white	-

KL Divergence Matrix for Image 2667



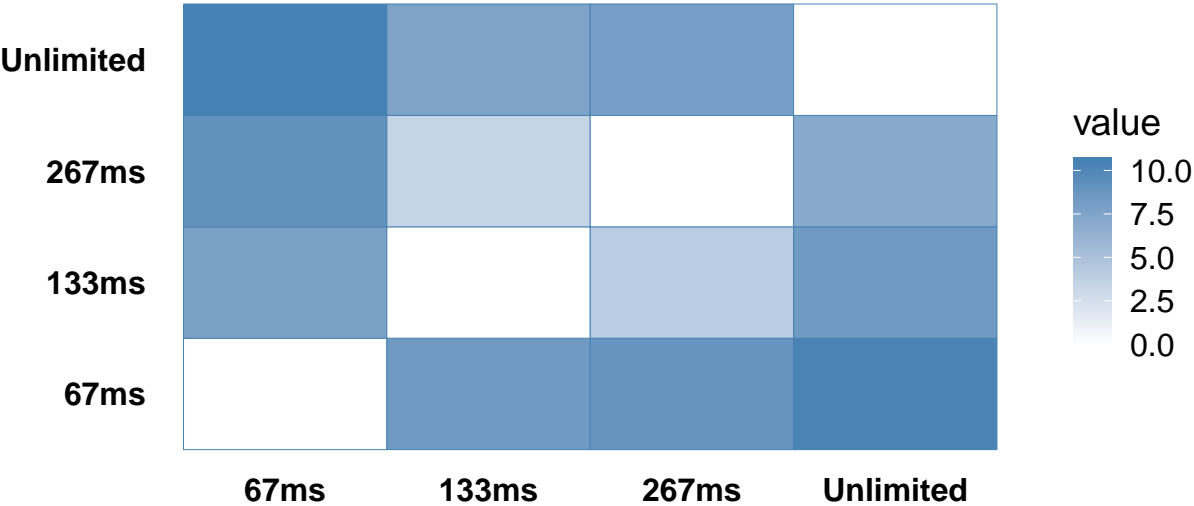
all_soas_alon	all_soas	67_alon	133_alon	267_alon
car	car	black	car	black
jeep	jeep	car	desert	car
-	road	desert	driving	jeep
-	truck	driving	jeep	tire
-	-	jeep	-	wheel
-	-	tire	-	-
-	-	wheel	-	-

KL Divergence Matrix for Image 2769



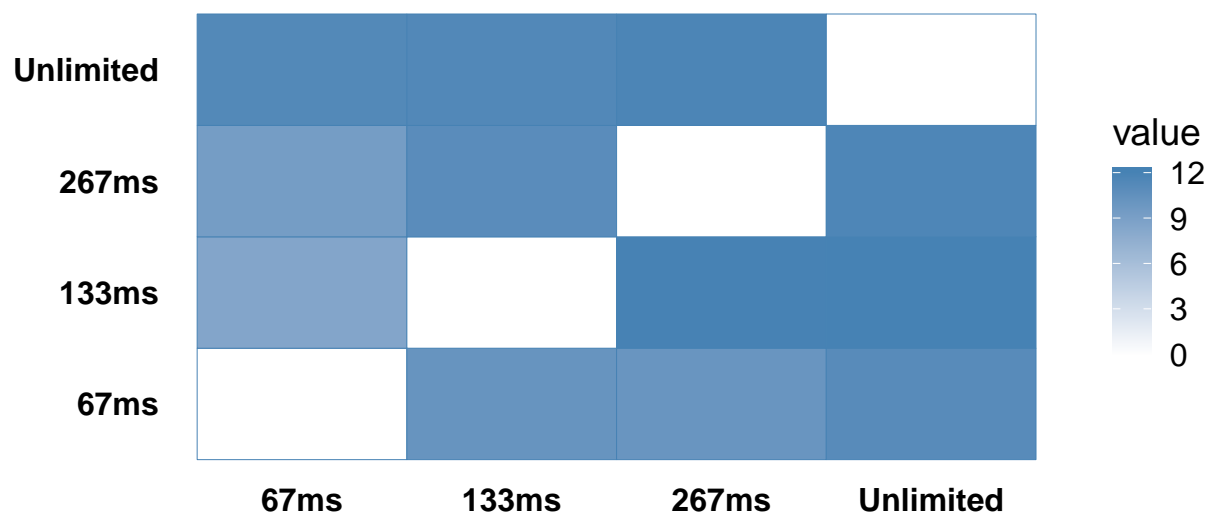
all_soas_alon	all_soas	67_alon	133_alon	267_alon
chalk	board	chalk	arm	chalk
-	chalk	green	chalk	map
-	hand	-	green	wrist
-	ring	-	man	-
-	write	-	teacher	-
-	-	-	writing	-

KL Divergence Matrix for Image 3279



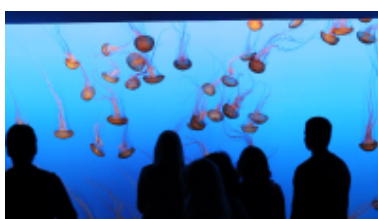
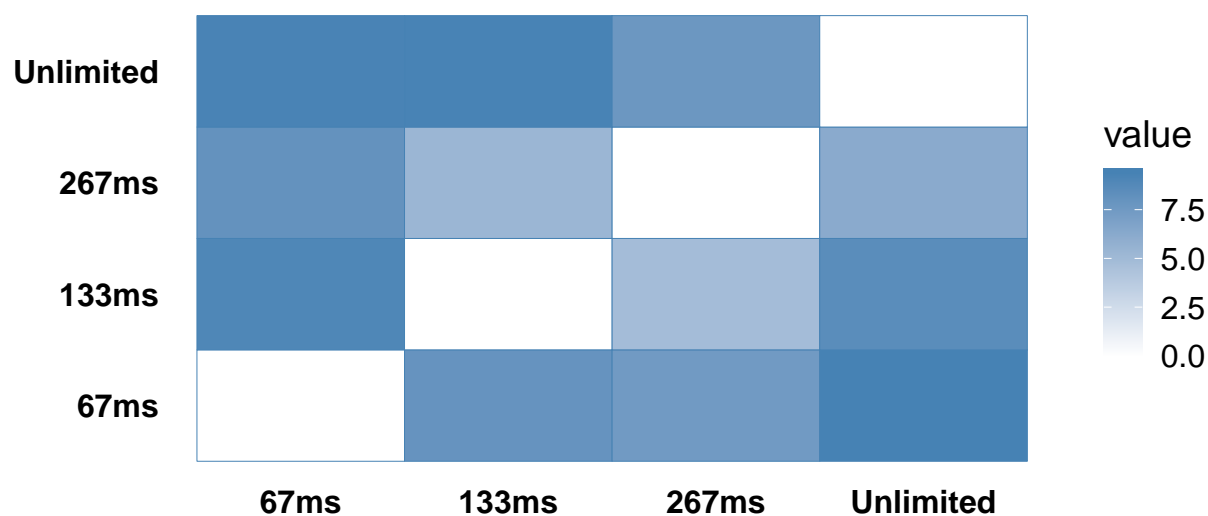
all_soas_alon	all_soas	67_alon	133_alon	267_alon
building	building	blue	building	building
sky	church	brown	hill	light
-	house	building	light	lighthouse
-	sky	sky	lighthouse	ocean
-	tree	-	ocean	sea
-	-	-	roof	sky
-	-	-	sea	-
-	-	-	sky	-

KL Divergence Matrix for Image 6806



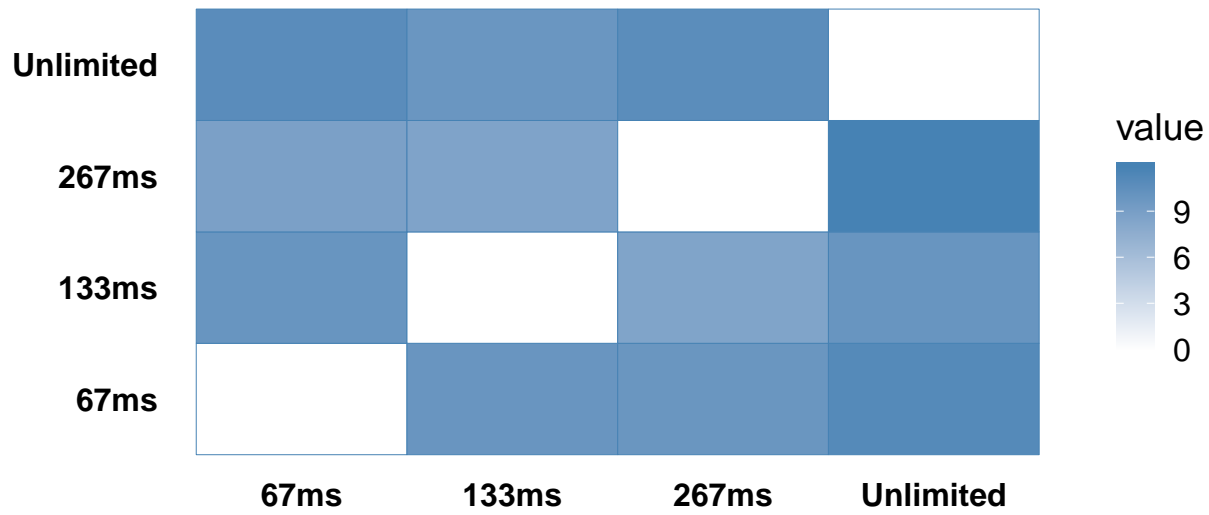
all_soas_alon	all_soas	67_alon	133_alon	267_alon
-	-	rope	rope	-

KL Divergence Matrix for Image 7360



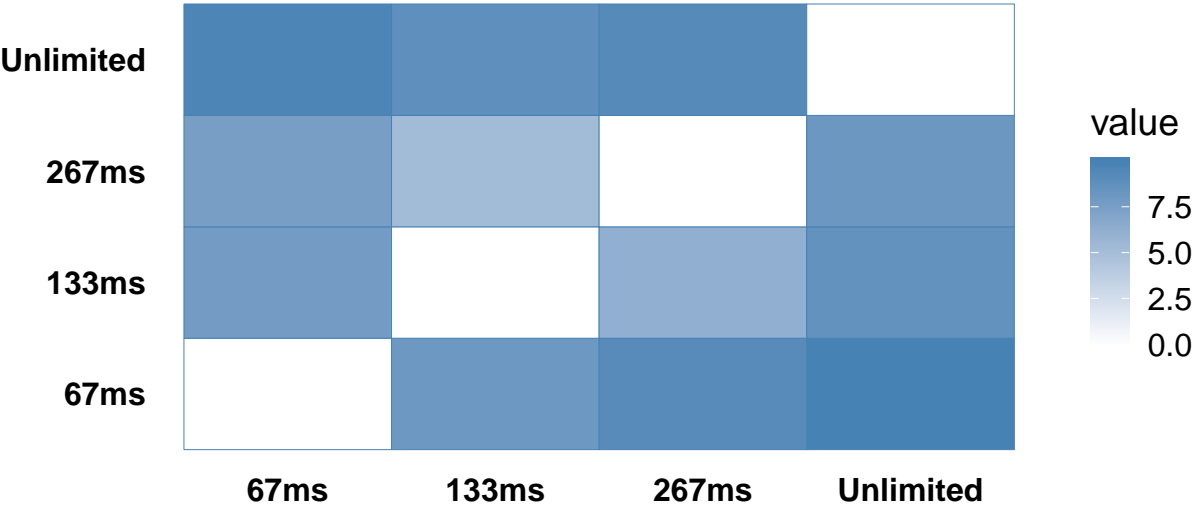
all_soas_alon	all_soas	67_alon	133_alon	267_alon
aquarium	aquarium	aquarium	aquarium	aquarium
people	man	light	huge	beautiful
water	people	moon	people	blue
-	water	people	tank	jellyfish
-	women	shadow	water	light
-	-	water	-	people
-	-	-	-	person
-	-	-	-	sea
-	-	-	-	tank
-	-	-	-	water

KL Divergence Matrix for Image 8858



all_soas_alon	all_soas	67_alon	133_alon	267_alon
food	food	food	food	food
-	red	-	many	-
-	-	-	white	-

KL Divergence Matrix for Image 8876



all_soas_alon	all_soas	67_alon	133_alon	267_alon
grass	forest	animal	animal	grass
lion	grass	grass	big	lion
-	lion	lion	grass	meat
-	-	watching	lion	tiger
-	-	-	meat	-
-	-	-	wild	-